Sciences of Europe

No 195 (2026)

Sciences of Europe
(Praha, Czech Republic)

ISSN 3162-2364
The journal is registered and published in Czech Republic.
Articles in all spheres of sciences are published in the journal.

Journal is published in Czech, English, Polish, Russian, Chinese, German and French, Ukrainian.
Articles are accepted each month.
Frequency: 24 issues per year.
Format - A4
All articles are reviewed
Free access to the electronic version of journal

Edition of journal does not carry responsibility for the materials published in a journal.
Sending the article to the editorial the author confirms it's uniqueness and takes full responsibility for possible consequences for breaking copyright laws.

Chief editor: Petr Bohacek
Managing editor: Michal Hudecek

- Jiří Pospíšil (Organic and Medicinal Chemistry) Zentiva
- Jaroslav Fähnrich (Organic Chemistry) Institute of Organic Chemistry and Biochemistry Academy of Sciences of the Czech Republic
- Rasa Boháček – Ph.D. člen Česká zemědělská univerzita v Praze
- Naumov Jaroslav S., MD, Ph.D., assistant professor of history of medicine and the social sciences and humanities. (Kiev, Ukraine)
- Viktor Pour – Ph.D. člen Univerzita Pardubice
- Petrenko Svyatoslav, PhD in geography, lecturer in social and economic geography. (Kharkov, Ukraine)
- Karel Schwaninger – Ph.D. člen Vysoká škola báňská – Technická univerzita Ostrava
- Václav Pittner -Ph.D. člen Technická univerzita v Liberci
- Dudnik Oleg Arturovich, Doctor of Physical and Mathematical Sciences, Professor, Department of Physical and Mathematical management methods. (Chernivtsi, Ukraine)
- Konovalov Artem Nikolaevich, Doctor of Psychology, Professor, Chair of General Psychology and Pedagogy. (Minsk, Belarus)

«Sciences of Europe» -
Editorial office: Křižíkova 384/101 Karlín, 186 00 Praha

E-mail: info@european-science.org
Web: www.european-science.org

CONTENT

ART STUDIES

Pogrebnyak M.
JOHN NEUMEIER'S DANCE THEATRE ON STAGE AND IN FILM: CREATIVE METHOD, FEATURES OF DIRECTING AND DANCE COMPONENT OF IS INDIVIDUAL PERFORMANCES 4

CHEMICAL SCIENCES

Klimko Yu., Koshchii I.
SYNTHESIS OF NOVEL ADAMANTYL-CONTAINING 5-AMINOPYRAZOLE INHIBITORS: EVALUATION OF THE IMPACT ON UBIQUITIN-SPECIFIC PROTEASE 7.....8

ECONOMIC SCIENCES

Amirov A.
SCREENING, SCALING, AND THE ALLOCATION FRONTIER: A MODEL OF INNOVATION-CAPITAL COMPOSITION AND LONG-RUN PRODUCTIVITY GROWTH 12

Asgarov A., Cahangirzada R.
FROM NUCLEAR SPECTRA TO PORTFOLIO RISK: RANDOM MATRIX THEORY AS A UNIFYING LANGUAGE FOR HIGH-DIMENSIONAL RISK ESTIMATION 24

Ibrayeva E., Ibrayev A., Robinson A.
CURRENT STATE AND DEVELOPMENT PROSPECTS OF KAZAKHSTAN'S DIGITAL ECONOMY 32

Kalinina E.
PREDICTIVE AI MODELS FOR REAL-TIME CONSUMER BEHAVIOR ANALYSIS IN MARKETING OPTIMIZATION 36

Muradov A.
RESOURCE WEALTH AND THE ALLOCATION OF ECONOMIC EFFORT: A THEORETICAL FRAMEWORK FOR STRUCTURAL TRANSFORMATION 43

GEOLOGICAL AND MINERALOGICAL SCIENCES

Boboyorov Kh., Rakhmatov A., Qodirov M., Mukhammadkulov N., Fazliddinova M.
LITHOLOGICAL-PETROGRAPHIC DESCRIPTION OF ROCKS OF BOKANTOV MOUNTAINS (KOKPATAS SUITE) 49

Bochkarev V., Bochkarev A., Bochkarev Ar.
THE S2 PARAMETER OF SHALE FORMATIONS: UNREALIZED HYDROCARBON POTENTIAL UNDER THE THERMODYNAMIC REGULATION OF FORMATION PRESSURE 54

MEDICAL SCIENCES

Fedorov M.
OPTICAL ARCHITECTURE AND THE BULK-VERSUS-SPATIAL CONFOUND IN HYBRID LED PHOTOPOLYMERIZATION OF NAIL COATINGS 67

PHILOLOGICAL SCIENCES

Allahverdiyeva A.
THE IMPACT OF MULTILINGUALISM ON SOCIAL INTEGRATION 72

SOCIAL SCIENCES

Aliyev U.
INTERNATIONAL EXPERIENCE IN THE IMPLEMENTATION OF STATE SOCIAL ORDERS FOR SOCIAL SERVICES BY NGOs.....76

TECHNICAL SCIENCES

Gavrilov A.
INTEGRATING MACHINE LEARNING INTO URBAN NETWORKS: EVALUATING ADVANCED INFRASTRUCTURE FRAMEWORKS.....85
Korviakov K.
A PREDICTIVE MULTI-SOURCE FRAMEWORK FOR CONTINUOUS POTABLE WATER SUPPLY IN MISSION-CRITICAL INFRASTRUCTURE: ATMOSPHERIC GENERATION, SOURCE ARBITRATION, AND QUALITY INVARIANCE 90
Herhi M., Kyryk Ya., Ratnikov D., Hrynchevskyi S.
NONLINEAR INTERDEPENDENCIES IN THE FORMATION OF INFRASTRUCTURE RESILIENCE: DAMAGE MODELING, FUNCTIONAL RECOVERY, AND POST-CRISIS MODERNIZATION.....96
Poliakov B., Pashko A.
AUTOMATION OF SOFTWARE DEVELOPMENT LIFECYCLE MANAGEMENT PROCESSES USING LLM AGENTS: INDUSTRIAL DEPLOYMENT EXPERIENCE ..104

ART STUDIES

ТАНЦТЕАТР ДЖОНА НОЙМАЙЄРА НА СЦЕНІ І В КІНО: ТВОРЧІЙ МЕТОД, ОСОБЛИВОСТІ РЕЖИСУРИ І ТАНЦЮВАЛЬНОЇ СКЛАДОВОЇ ЙОГО ОКРЕМИХ ВИСТАВ

Погребняк М.М.
Полтавський національний педагогічний університет імені В. Г. Короленка, доктор мистецтвознавства, професор

JOHN NEUMEIER'S DANCE THEATRE ON STAGE AND IN FILM: CREATIVE METHOD, FEATURES OF DIRECTING AND DANCE COMPONENT OF IS INDIVIDUAL PERFORMANCES

Pogrebnyak M.
Poltava National Pedagogical University named after V. G. Korolenko, Doctor of Stady of Art, Professor
DOI: 10.5281/zenodo.22107020

АНОТАЦІЯ

У статті проаналізовано найвідоміші авторські творчі роботи на сцені і в кіно одного з представників сучасного балету XXI ст. – балетмейстера Джона Ноймайєра. Виявлені і систематизовані, *по-перше, особливості творчого методу* балетмейстера, а саме: 1) створення балетів великої форми, лібретто більшості з яких пов'язані з шедеврами світової літератури; а балетмейстер одночасно є декоратором, художником по костюмам та з освітлення; 2) використання так званої «небалетної музики»; 3) оригінальні прийоми монтажу сценарного матеріалу та вирішення мізансцен і сценографії з активним залученням масового кордебалету. *По-друге, особливості танцювальної складової його вистав*: використання виразних засобів різних стильових напрямів театрального танцю, пантоміми, прийомів контактної імпровізації, акторської гри для створення різноманітних художніх образів; створення дуету-діалогу з використанням нестандартних пластичних рішень. Наведені приклади режисури вистав балетмейстера, зокрема використання таких *режисерських прийомів як «театр в театрі», «стоп кадр»;* прийомів в композиції танцю – *антиунісон, асиметрія, паралельність розгортання подій, контрапункт*.

ABSTRACT

Relevance of research. Dance theatre as an art form led to the final liberation of dance theatre from it's institutional connection with opera at the beginning of the 20th century. Dance theatre of the 21th centure uses all the ingredients of the theatrical production: the introduction of texts that are spoken, read or recited from off-stage; giving meaning to choreography, things, costumes, careful coordination of all stage materials. However, the need for theoretical support for artists and the educational process in this field, the study of creative methods and new ideas of ballet masters – the most famous representatives of world modern ballet – still remains a pressing problem.

The purpose of this scientific exploration is to identify and systematize the features of the creative method, of directing and dance component of performance on stage and in cinema by choreographer John Neumeier.

The research methodology is based on the using of the biographical, source studies and the method of included observation in the study of John Neumeier's creative works on stage and cinema; stylistic analysis of ballets and synthesis to identify and systematize the features of the choreographer's creative method; the dance component and directorial techniques of his performances.

The scientific novelty lies in the identifying and systematizing the features of the creative method, of directing and the dance component of performances on stage and in cinema by choreographer John Neumeier.

Results. As a result of the analysis of the most famous creative works of the author on stage and in the cinema, one of the representatives of contemporary ballet of the 21st century, choreographer John Neumeier, the following features of the choreographer's creative method were identified and systematized: 1) creation of Large-scale ballets, the librettos of most of which are related to masterpieces of world literature, and the ballet master is simultaneously a set designer, costume designer and lighting designer; 2) the using of so-called non-ballet music; 3) original techniques for editing script material and solving mise-en-scene and scenography with active involvement of the mass corps-de-ballet. Secondly, the features of the dance component of his performances: the using of expressive means of various stylistic directions of theatrical dance, pantomime, contact improvisation techniques, acting to create various artistic image; the creation of a duet-dialogue using non-standard plastic solutions. Examples of the choreographer's directing of performances are given, in particular the using of such directing techniques in dance composition – anti-unison, asymmetry, parallel development of events, counterpoint.

Ключові слова: танцтеатр, сучасний театральний танець, балетмейстер, творчий метод, режисура, композиція танцю.

Keywords: dance theatre, contemporary theatrical dance, choreographer, creative method, direction, dance composition.

Постановка проблеми. Минуло вже більш ніж 100 років з моменту виникнення явища сучасного авторського танцтеатру. Але, й досі залишається актуальною необхідність теоретичної підтримки митців і навчального процесу у галузі нових явищ хореографічного мистецтва, нових ідей хореографів, зокрема одного з найвідоміших представників світового сучасного балету Джона Ноймайєра. Проблемам теорії та історії сучасного танцтеатра і нових напрямів театрального танцю XX–XXI ст. присвячені нечисельні праці західних дослідників: Джека Андерсон, Дон Макдонах, Ендрю Марк Вентинк, Агнесса де Міль, Ізі Парш-Бергсон, Еврістін Стодел, Маринелла Гваттеріні та інших. Але поза межами їхніх наукових розвідок залишається творчість Джона Ноймайєра. Український вчений Олександр Чепалов на сторінках своєї монографії торкається драматургійної складової балетів Джона Ноймайєра [1, с. 221–248]; а Марина Погребняк робить *порівняльний аналіз авторських версій хореографічної драми* «Дама з камеліями» у постановці Дж. Ноймайєра і Аніко Рехвіашвілі [2, с. 253–254]. Виявлення і систематизація особливостей творчого методу, режисури і танцювальної складової вистав на сцені і в кіно балетмейстера Джона Ноймайєра залишається поза межами зазначених наукових розвідок, що і є **метою** даної статті. **Методологія дослідження.** Авторкою використано такі **методи дослідження**: *джерелознавчий; біографічний, включеного спостереження* у вивченні творчих робіт Джона Ноймайєра на сцені і в кіно; *стильового аналізу балетів та синтезу* – для виявлення і систематизації особливостей творчого методу балетмейстера; танцювальної складової та режисерських прийомів його вистав.

Виклад основного матеріалу. Джон Ноймайєр, балетмейстер Франкфурктського балету, відшукував героїв своїх вистав у світовій літературі, міфології, музиці. За його словами, він намагався «активно експериментувати у напрямку так званого «симфонічного танцю», який дозволяє знайти засіб розповідати історії, близькі до сучасного життя, до молоді». Загалом він спробував утримати жанр «великого балету» на плаву [1, с. 222].

«Танцтеатр», як мистецька форма, зумовила на початку XX ст. остаточне вивільнення театру танцю з інституціального зв'язку з оперою. Як зазначає П. Паві, «у танцювальному театрі на відміну від звичайного театру, який використовує танець, рухи та хореографію, танець «залучається як театральний ефект» … Танцювальний театр використовує всі інгредієнти театральної постанови: введення текстів, які виголошують, читають або вимовляють з-поза сцени; надання значення хореографії, речам, костюмам, ретельна координація всіх сценічних матеріалів» [3, с. 511].

У творчій спадщині балетмейстера: авторські версії відомих класичних балетів («Лебедине озеро», «Спляча красуня», «Жизель», «Лускунчик», «Попелюшка», «Дафніс і Хлоя»); танцсимфонії на музику Г. Малера («Третя симфонія» Малера; «Шоста симфонія» Г. Малера; «Єпілог»); сучасні сюжетні балети, зокрема за мотивами відомих драматичних творів («Дама з камеліями», «Ромео і Джульєтта», «Отелло», «Дон Жуан», «Пер Гюнт», «Русалочка», «Трамвай "Бажання"» та ін.); вистави на релігійні теми та духовну музику («Реквієм» В. А. Моцарта, «Легенда про Йосипа» Р. Штрауса, «Месія» Генделя-П'ярта, «Страсті за Матфеєм» та ін.). У сфері інтересів хореографа – весь історичний діапазон світової літератури: від біблії, античної трагедії та середньовічних саг до сучасної драматургії та поезії [1, с. 223].

Відповідно набувають популярності також фільми-балети Джона Ноймайєра. Фільм-балет – особливий жанр кінематографічного мистецтва, що комбінує художні засоби кіно з художніми засобами власне балету. Фільм-балет є оригінальною кінорежисерською роботою, здійсненою павільйонними, або натурними зйомками із застосуванням різноманітних кінооператорських технік, прийомів монтажа і таке ін. з можливою участю драматичних акторів. У фільмографії Джона Ноймайєра можна побачити фільми: «Дама з камеліями» (Die Kameliendame, Гамбурзький балет, 1987-й р.); «Легенда про Йосипа» (Josephs Legende, Гамбурзький балет, 1991 р.); «Ілюзії як "Лебедине озеро"» (Illusions like «Swan Lake», Гамбурзький балет, 2001 р.); «Смерть у Венеції» (Der Tod in Venedig, Гамбурзький балет, 2003-й р.); «Страсті за Матфеєм» (Matthäus-Passion, Гамбурзький балет, 2005-й р.); «Сильвія» (Паризька національна опера, 2006-й р.); «Дама з камеліями» (La Dame aux Camelias, Паризька національна опера, 2008-й р.); «Русалочка» (Балет Сан-Франциско, 2011-й р.); «Різдвяна ораторія» Баха (2015-й р.); «Проєкт Бетховен» (2018-й р.); «Скляний звіринець» (2019-й р.).

Виявлено, що *особливостями творчого методу* Джона Ноймайєра є те, що чисельні його лібретто пов'язані з шедеврами світової літератури Евріпіда, Мольєра, О. Дюма, Г. Ібсена, Т. Манна та інших. *Головним у виборі літературного твору,* за словами балетмейстера, є наявність поміж рядків того, *що можна виразити танцем*. А головним у виборі музиці – та, що пробирає до глибини душі, щоб виникло бажання встати і рухатися разом з нею. Дж. Ноймайєр чує сюжет в музиці. Для нього головна відправна точка для створення художнього образу – емоція. Балетмейстер сам *складає музичну партитуру своїх вистав* (музика сучасних класиків, романтиків, авангардистів); найчастіше він *сам є сценографом, художником по костюмам і світлу*, знаходячі для кожного сюжету особливі фарби. За його словами, «... митець повинен мати своє особливе відчуття світу; чути свою особливу музику» [4]. Джон Ноймайєр впевнений, що у мистецтві є головним те, що ми відчуваємо, а не те що ми конструюємо за допомогою логічних висновків. І досить важлива для нього правдивість того, про що він говорить [4].

Розглянемо *особливості режисури і танцювальної складової* на прикладі окремих вистав балетмейстера.

Постановку Дж. Ноймайєра *«Дама з камеліями»* (2008-й р.) на музику Ф. Шопена можна

умовно назвати «дуетною виставою» за переважання в ній дуетної форми класичного і неокласичного танцю головних героїв і кордебалету.

В першому акті – дует зустрічі Маргарити і Армана; у другому акті – любовний дует; у третьому акті – дует спочатку «сварки», а потім пристрасних стосунків, вирішені засобами неокласичного танцю з використанням трансформованих підтримок у «падаючі пози», партерних і повітряних підтримок, обводок з «вільним» корпусом і «вільною» пластикою рук.

Трагічна доля куртизанки XIX ст. протягом усього балету перекликається з долею куртизанки XVIII ст. Манон, хореографічний текст, що розкриває її художній образ, і образ її численних залицяльників побудований на лексиці класичного танцю.

У композиції цієї вистави Дж. Ноймайєр використовує виразні засоби танцтеатру. А саме: *пластичну гру акторів* у сцені аукціону, влаштованого у квартирі померлої куртизанки Маргарити Готьє, на якому розпродаються меблі і особисті речі хазяйки, а також *у сцені візиту* месьє Дюваля до Маргарити; *режисерський прийом «театр в театрі»* в сцені присутності Маргарити на виставі балету «Манон Леско», у якому відома куртизанка XVIII ст. обманює Де Гріє з численними залицяльниками (ця історія є лейтмотивом усього спектаклю). У побудові мізансцен і композиції танцю: *прийом «стоп кадру»*, коли головні герої Маргарита і Арман поєднуються в «дзеркальній позі» з героями балету «Манон Леско», що посилює асоціації глядачів стосовно перенесення долі танцівників на головних героїв; *прийом асиметрії і антиунісону під час одночасності розгортання подій; паралельного монтажу сценарного матеріалу* (Маргарита танцює у парі на фоні діючого танцювального ансамблю або Герцог шукає Маргариту на фоні ансамблю танцюючих на балу) і *контрастного* монтажу (Прюданс і Гастон у живому романтичному танці показують захопленість один одним, а Маргарита і Арман ведуть спокійну бесіду) [5; 2, с. 253].

У *«Отелло»* балетмейстер будує поліфонічну структуру з великою кількістю пластичних характеристик методом паралельного монтажу. Використовує різні стильові напрями танцю. Наприклад, «шляхетні венеціанці» розмовляють мовою класичного танцю; також у цьому стилі Отелло дає відповідь сенату. Поряд існує історико-побутовий танець мореска; пластична акторська гра в уповільненому, розтягнутому темпі – в Адажіо Отелло і Дездемони. Цікавий режисерський прийом «поетичного двійника» і прийом контактної імпровізації використовує Дж. Ноймайєр для візуалізації внутрішнього образу головних героїв. Так, разом з Дездемоною виникає її двійник – Примавера у оточенні грацій, яка розкриває свої почуття «вільним» імпресіоністичним танцем; а поряд зі шляхетним Мавром – його двійник екзотичний дикун, пристрастна хореографія якого побудована засобами експресіоністичного танцю з виразними стрибками, обертами, джазовими підскоками і кроками [6].

«Страсті за Матфеєм» на музику І.-С. Баха – балет-симфонія розігрується в атріумі собору і оповідає про події Страстного тижня та Воскресіння Христового. Вистава має свою *оригінальну сценографію:* невеликий задрапований поміст – Голгофу. Довгі лави перетворюються на зловісні ланцюги для Христа. Оголеність тіла Христа символізує його істинну людську природу, здатну до страждання заради спасіння людства. Хореографічний текст побудований з використанням *побутової пластики* (штовханням Христа та ін.); *пластичної акторської гри* (розп'яття Христа, тощо); *експресіоністичного танцю.* У танці кордебалета балетмейстер використовує прийом *асиметрії і антиунісону*. Сценографію доповнюють статичні фігури артистів кордебалету, які виразно розміщені на сходинках храму і періодично вступають в рухому дію [7].

«Проєкт Бетховен» – створений в ювілейний рік композитора. Перші сцени вистави присвячені біографії Л. Бетховена; частина між першою і другою має назву «Інтермеццо» на музику Л. Бетховена до балета «Прометей», а в другій половині вистави «Героїчна» – хореографічні сцени на музику Третьої симфонії композитора. В цій виставі Джон Ноймайєр є балетмейстером, художником по костюмам та з освітлення, декоратором. *Хореографія дуетної форми* (контактної і безконтактної) побудована виразними засобами *неокласичного танцю. Сценографію доповнює рухома пластика та акторська гра кардебалету за принципом контрапункту* [8].

Оскільки кліпове мислення стало сьогодні популярним, окремі сцени балетмейстер робить свідомо довгими. Наприклад, в першій сцені «Русалочки» для балетмейстера дуже важливо, щоб глядачі проникли в той світ, де зміниться русалочка. Для Дж. Ноймайєра ця казка про її автора Х. Андерсона. Сенс задуму вистави Ноймайєра полягає в тому, що у фіналі починається новий етап, а саме творець і його творіння продовжують життя у Вічності.

Висновок. Таким чином, з аналізу балетів на сцені і в кіно Джона Ноймайєра виявлено: *по-перше* – особливостями творчого методу балетмейстера є: 1) створення балетів великої форми, лібретто більшості з яких пов'язані з шедеврами світової літератури; а балетмейстер одночасно є декоратором, художником по костюмам та з освітлення; 2) використання так званої «небалетної музики»; 3) оригінальні прийоми монтажу сценарного матеріалу та вирішення мізансцен і сценографії з активним залученням масового кордебалету. *По-друге, особливості танцювальної складової його вистав*: використання виразних засобів різних стильових напрямів театрального танцю, пантоміми, прийомів контактної імпровізації, акторської гри для створення різноманітних художніх образів; створення дуету-діалогу з використанням нестандартних пластичних рішень. Прикладами режисури вистав балетмейстера є: використання таких *режисерських прийомів як «театр в театрі», «стоп кадр»;* прийомів в композиції танцю – *антиунісон, асиметрія, паралельність розгортання подій, контрапункт*.

References

1. Чепалов О. Хореографічний театр Західної Європи XX ст.: монографія. Харків: ХДАК, 2007. 344 с.

2. Погребняк М. М. Нові напрями театрального танцю XX – початку XXI ст.: історико-культурні передумови, крос-культурні зв'язки, стильова типологія: монографія. Полтава: ТОВ «Прінтмакс», 2021. 330 с.

3. Паві П. Театр (Танцювальний). *Словник театру* / пер. з франц. М. Якубяк. Львів: ЛНУ імені Івана Франка, 2006. С. 510–512.

4. Абсолютний слух. Джон Ноймайер: веб-сайт. URL: https://youtu.be/Kaq3BEZ_k1k?si=hxrKsD2EzGbRZ3yd (дата звернення 5.01.2020).

5. The Lady of the Camellias: веб-сайт. URL: https://youtu.be/u29kfBhTCY0?si=SJHZzcn271nSN18D (дата звернення 5.01.2020).

6. Отелло: веб-сайт. URL: https://youtu.be/0cxQyWy-K7w?si=q742sCat5rA0f7H3 (дата звернення 17.03.2026).

7. I S Bach Matthaus Passion BWV 244 Ballet von John Neumeier Teil 2: веб-сайт. URL: https://youtu.be/4YzABkNBf5M?si=PxeqR9ShLdUuGUfN (дата звернення 16.03.2026).

8. Beethoven Project – Ballet by John Neumeier: веб-сайт. URL: https://youtu.be/WXTIPTKUvus?si=rbUWuGIaeljNTmCS (дата звернення 15.03.2026).

CHEMICAL SCIENCES

SYNTHESIS OF NOVEL ADAMANTYL-CONTAINING 5-AMINOPYRAZOLE INHIBITORS: EVALUATION OF THE IMPACT ON UBIQUITIN-SPECIFIC PROTEASE 7

Klimko Yu.
PhD in Chemistry, Senior Lecturer
Department of Organic Chemistry and Technology of Organic Substances
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"
Koshchii I.
PhD in Chemistry, Associate Professor
Department of Organic Chemistry and Technology of Organic Substances
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"
DOI: 10.5281/zenodo.22107034

ABSTRACT

To further expand the structure-activity relationships (SAR) of the previously published ubiquitin-specific protease 7 (USP-7) inhibitor STIRUR-41, adamantyl-containing 5-aminopyrazolones (1) and (2) were designed and synthesized. The chemical identity of the desired structure was confirmed by nuclear magnetic resonance. All new derivatives were tested as potential USP-7 inhibitors. Compounds 1a–d inhibited enzymatic activity in a dose-dependent manner and with lower IC_{50} values compared to the parent compound STIRUR-41. Notably, compound (1), containing an adamantyl group linked to a carbamate moiety, showed itself to be the most active candidate. In contrast, compounds belonging to series (2), which exhibited high steric hindrance, showed no activity. Compounds from series (1) are not cytotoxic to a number of tumor and normal cell lines at a concentration of 10 μM.

Keywords: antiproliferative agents, pyrazole compounds, ubiquitinspecific protease 7 inhibitors, 4-(1-adamantyl)phenyl isocyanate.

Introduction

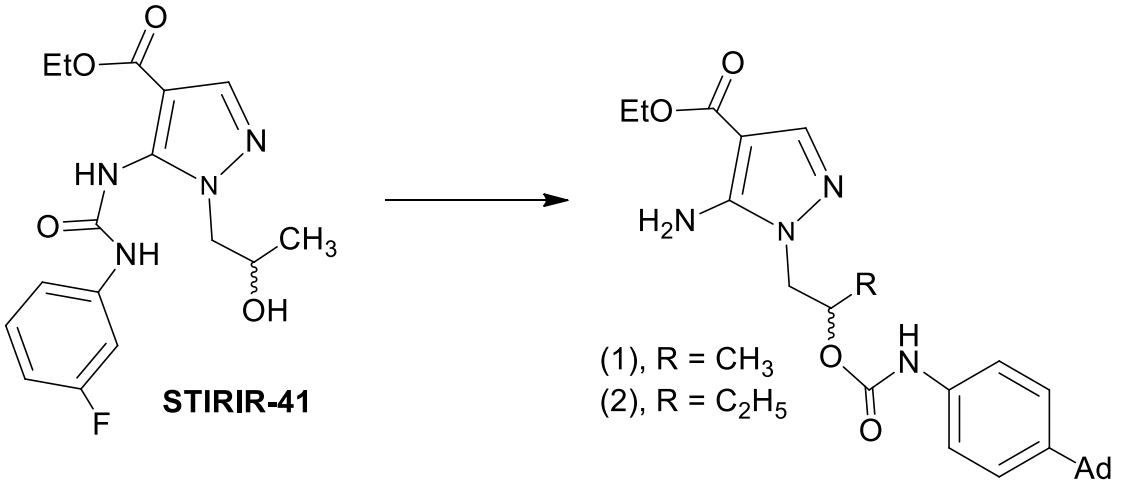
The ubiquitin-proteasome system is a major protein regulatory system present in all eukaryotic cells, allowing the regulation and therefore removal of all misfolded, aggregated or unnecessary proteins, thereby maintaining cellular homeostasis [1].

Protein ubiquitination is a post-translational dynamic and a reversible process that can be reversed by deubiquitinating enzymes called deubiquitinases (DUBs). Of the seven DBU families, the ubiquitin-specific protease 7 (USP-7) family (also known as herpes virus-associated ubiquitin-specific proteases) is the largest [2]. In the human genome, the USP-7 family consists of 58 members that regulate many cellular proteins, including those involved in viral replication, immune response, epigenetic control, apoptosis, DNA damage response, as well as DNA replication and transcription. The role of USP-7 in the regulation of these cellular pathways suggests that this enzyme is a promising new therapeutic target for oncological therapy [3,4]. Notably, increased USP-7 activity correlates with the occurrence and subsequent progression of various human tumors through interactions with oncogenic proteins such as MDM2, MDMX [5], β-catenin [6], NF-κB [7], FOXP3 [8], and the tumor suppressors p53 and PTEN [9]. Furthermore, USP-7 activity is also involved in cancer angiogenesis, promoting the formation of new blood vessels in the tumor microenvironment [10]. Its role in tumor pathology is twofold: it acts as both a promoter and a tumor suppressor, depending on the cancer type [11,12]. Specifically, overexpression of USP-7 has been observed in many malignancies, where it serves as a prognostic marker for cancer progression and poor prognosis. Conversely, low USP-7 expression has been detected in some neuroendocrine lung tumors [13] and in non-small cell lung cancer, adenocarcinoma [14]. This suggests that the role of USP-7 may vary depending on the cancer type, those involved in viral replication, immune response, epigenetic control, apoptosis, DNA damage response, as well as DNA replication and transcription. The role of USP-7 in the regulation of these cellular pathways suggests that this enzyme is a promising new therapeutic target for oncological therapy [3,4]. Notably, increased USP-7 activity correlates with the occurrence and subsequent progression of various human tumors through interactions with oncogenic proteins such as MDM2, MDMX [5], β-catenin [6], NF-κB [7], FOXP3 [8], and the tumor suppressors p53 and PTEN [9]. Furthermore, USP-7 activity is also involved in cancer angiogenesis, promoting the formation of new blood vessels in the tumor microenvironment [10]. Its role in tumor pathology is twofold: it acts as both a promoter and a tumor suppressor, depending on the cancer type [11,12]. Specifically, overexpression of USP-7 has been observed in many malignancies, where it serves as a prognostic marker for cancer progression and poor prognosis. Conversely, low USP-7 expression has been detected in some neuroendocrine lung tumors [13] and in non-small cell lung cancer, adenocarcinoma [14]. This suggests that the role of USP-7 may vary depending on the cancer type.

For all these reasons, in recent years, USP-7 has emerged as a new and promising therapeutic target for cancer treatment, and several covalent and non-covalent small molecule inhibitors have been developed, showing promising efficacy in preclinical studies [15–23]. In particular, monocyclic (e.g., thiazoles and pyridines) or bicyclic (e.g., quinazolines, pyrazolopyrimidines, thienopyrimidines, and recently published and patented imidazotriazines and tricyclic derivatives (e.g., triterpenoid derivatives) [16].

In this context, previous studies have led to the identification of a large library of pyrazole derivatives capable of blocking IL-8-induced chemotaxis of human neutrophils [24]. In particular, the most active compound, STIRUR-41 (Fig. 1), was subjected to additional studies, showing significant inhibition of the migratory capacity of HTLA-230 neuroblastoma cells [25].

Subsequent drug-sensitive target stability studies, combined with limited proteolysis, as well as in vitro and in cell studies, demonstrated the ability of STIRUR-41 to inhibit both the enzymatic activity of USP-7 and its expression levels in neuroblastoma-associated cells [26].



Scheme 1. Molecular structure of the compound STIRUR-41 and new 5-aminocarbamoyl derivatives (1) and (2).

To further expand the structure–activity relationship (SAR) of STIRUR-41, 5-aminopyrazoles (5APs) (1) and (2) were developed (Scheme 1). Specifically, the phenylaminocarbonyl moiety was formally transferred from the 5-nitrogen atom of STIRUR-41 to the side-chain oxygen atom, resulting in carbamate analogs (1) and (2). The structure of the new compounds was refined by introducing an adamantyl substituent on the carbamate-linked phenyl ring and by altering the side-chain length (i.e., the 2-hydroxypropyl and 2-hydroxybutyl moieties).

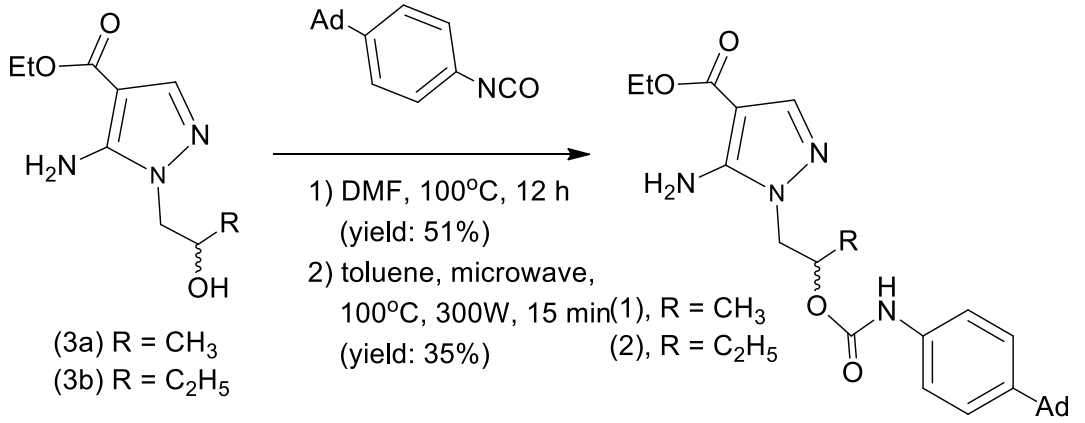
The physiological activity of framework hydrocarbon derivatives, primarily adamantane derivatives, has been extensively studied. The activity of these compounds is due to the pronounced lipophilic nature of the compact framework hydrocarbon fragment. Therefore, the aim of this study was to attempt to impart greater lipophilicity to known anti-priposomal drugs, which, in our opinion, could improve their pharmacological activity.

We confirmed the formation of the desired structures using nuclear magnetic resonance (NMR).

The physiological activity of framework hydrocarbon derivatives, primarily adamantane derivatives, has been extensively studied. The activity of these compounds is due to the pronounced lipophilic nature of the compact framework hydrocarbon fragment. Therefore, the aim of this study was to attempt to impart greater lipophilicity to known anti-priposomal drugs, which, in our opinion, could improve their pharmacological activity.

Results and Discussion

Compounds (1) and (2) were prepared starting from 5-aminopyrazole intermediates (3a,b), synthesized as reported previously (Scheme 2) [24]. Condensation of (3a) with an adamantyl-substituted isocyanate in anhydrous DMF at 100 °C for 12 h resulted in the isolation of compounds 1a–d in moderate yields of 51%.



Scheme 2. Synthesis of compounds (1) and (2). 1) Adamantyl-substituted phenyl isocyanate, DMF, 100 °C, 12 h (yield: 51%) and (2) excess of the same phenyl isocyanate, toluene, microwave, 100 °C, 300 W, 15 min (yield: 35%).

Condensation of (3a) with an adamantyl-substituted isocyanate in anhydrous DMF at 100 °C for 12 h afforded compound (1) in a moderate yield of 51%. Unfortunately, under these reaction conditions, 5-aminopyrazole (3b) proved unreactive toward isocyanates, possibly due to increased steric hindrance of the ethyl group adjacent to the alcohol functional group. Therefore, compound (2) was prepared via an alternative synthetic procedure involving the use of anhydrous toluene as solvent, excess isocyanate reagent (3 equivalents), and two microwave heating cycles (300 W, 7 min). With both methods, the final compounds were isolated by column chromatography (silica gel) using dichloromethane and ethyl ether as eluents.

Interestingly, pyrazoles (3) contain two nucleophilic centers (i.e., NH_2 and OH groups) potentially capable of reacting with isocyanate reagents. However, different reaction conditions were found to affect the regioselectivity of the condensation.

In more detail, as previously reported, the use of a non-polar solvent (toluene) under normal heating conditions oriented the reaction towards the amino group of compound (3), which led to the release of urea derivatives [24,27]. In contrast, heating the toluene mixture under microwave irradiation or using polar aprotic solvents (DMF) resulted in preferential reaction of the alcohol functional group, allowing the formation of carbamate compounds [28].

The chemical identity of compounds (1) and (2) was determined using NMR spectroscopy. In particular, the proton signals in NMR at 6.00 ppm. (also present in the spectra of the precursors (3) and the absence of a signal at 4.93 ppm. (assigned to OH in spectra (3) indicate the presence of a free amino group at position 5 and confirm functionalization of the hydroxyl group in derivatives (1) and (2).

Experimental

All reagents were purchased from Sigma-Aldrich. DMF was reagent grade and dried over molecular sieves (5 Å 1/16 inch beads). Unless otherwise stated, all commercial reagents were used without further purification. Organic solutions were dried over anhydrous sodium sulfate. Thin-layer chromatography using silica gel plates on aluminum backing (Merck DC-Alufolien Kieselgel 60 F254) was used for routine monitoring of reaction progress and confirmation of the purity of analytical samples. Pure dichloromethane (DCM) or a DCM/2% methanol mixture was used as the eluent, and spot detection was accomplished using UV light or iodine vapor. Merck silica gel, 230–400 mesh, was used for chromatography. Flash chromatography was performed on an Isolera One instrument (Biotage, Uppsala, Sweden) using silica gel columns. Melting points were determined on a BÜCHI M-560 (Buchi instruments, Flawil, Switzerland) and were uncorrected. Microwave oven: CEM Discover (CEM Corporation, Matthews, NC, USA), a single-mode microwave oven with a maximum output power of 300 W, fiber-optic temperature control, magnetic sample stirring, and an air-cooling system. ^1H and ^{13}C NMR spectra were recorded on a JEOL JNMECZR (400 MHz, Tokyo, Japan). Chemical shifts are reported in δ units (ppm) relative to the internal standard tetramethylsilane, and splittings are described as follows: s (singlet), bs (broad singlet), d (doublet), t (triplet), q (quartet), and m (multiplet). First-order values for the spin–spin coupling constants J are given in Hz.

High-resolution mass spectrometry (HRMS) data were obtained on an FTICR-MS instrument equipped with a q-TOF mass analyzer (Ionspec 7.0 T). Compounds (3a,b) were prepared as previously reported [24].

General Synthetic Procedure for the Preparation of Compound (1)

To a solution of (3a) (1 mmol) in DMF (5 ml) 4-(1-adamantyl)phenyl isocyanate (1.1 mmol) was added dropwise with stirring at room temperature. The mixture was stirred at 100 °C for 12 h, then cooled to room temperature and water (30 ml) was added. After basification with aqueous ammonia, the mixture was extracted with ethyl acetate (3 x 10 ml). The combined organic phases were washed with water (3 x 10 ml) and brine (3 x 10 ml), dried ($MgSO_4$) and evaporated under reduced pressure. The crude product was purified by flash chromatography (silica gel, eluent: DCM - DCM/50% Et_2O).

Methyl 5-Amino-1-(2-(((4-(1-Adamantyl)phenyl)Carbamoyl)Oxy)Propyl)-1H-Pyrazole-4-Carboxylate (1). Yield: 51%. White solid. Mp. 189–190 °C (DCM/Et_2O). ^1H NMR (400 MHz, DMSO-d_6, δ): 1.20–1.25 (m, 6H, $CHCH_3$+ CH3 ester), 1.48, 1.68, 1.92 (15H, Ad), 4.08–4.18 (m, 4H, OCH_2 + NCH_2), 5.09–5.15 (m, 1H, CHO), 6.28 (bs, 2H, NH_2, exchangeable), 7.01–7.12 (m, 2H, arom. H), 7.40–7.47 (m, 3H, arom. H + H pyraz.), and 9.60 (bs, 1H, NH, exchangeable). ^{13}C NMR (101 MHz, DMSO-d6) δ 163.96, 153.39, 150.87, 139.25, 135.95, 120.45, 115.89, 115.67, 94.24, 71.6, 69.14, 59.18, 51.04, 45.3, 37.5, 29.0, 18.16, and 15.04. MS (m/z): $M^{\bullet+}$ (450).

General Synthetic Procedure for the Preparation of Compound (2)

To a solution of (3b) (1 mmol) in toluene (5 ml) 4-(1-adamanti)phenyl isocyanate (1.5 mmol) was added dropwise with stirring at room temperature. The mixture was stirred under microwave irradiation (maximum power 300 W) at 90 °C for 7 min. Then another equivalent of phenyl isocyanate (1.5 mmol) was added, and the solution was again heated under microwave irradiation at 90 °C for 7 min. After cooling to room temperature, the solvent was evaporated under reduced pressure, and the crude product was purified by flash chromatography (silica gel, eluent: DCM - DCM/50% Et_2O).

Ethyl 5-Amino-1-(2-(((4-(1-Adamantyl)phenyl)Carbamoyl)Oxy)Butyl)-1H-Pyrazole-4-Carboxylate (2). Yield: 35%. White solid. Mp. 146–148 °C (DCM/Et_2O). ^1H NMR (400 MHz, DMSO-d6, δ): 0.91 (t, J = 7.3 Hz, 3H, CH_3), 1.23 (t, J = 7.1 Hz, 3H, CH_3 ester), 1.48–1.68 (m, 2H, $CHCH_2$), 1.48, 1.68, 1.92 (15H, Ad), 4.09–4.17 (m, 4H, OCH_2 +NCH2), 4.96–5.03 (m, 1H, CHO), 6.27 (bs, 2H, NH_2, exchangeable) 7.06–7.12 (m, 2H, arom. H), 7.39–7.45 (m, 3H, arom. H+H pyraz.), and 9.63 (bs, 1H, NH, exchangeable). ^{13}C NMR (101 MHz, DMSO-d6) δ 163.44, 153.14, 150.39, 138.68, 135.47, 119.89, 115.36,115.14, 93.75, 72.87, 71.6, 58.65, 49.17, 45.3, 37.5, 24.43, 29.0, 14.51 and 9.26. MS (*m/z*): $M^{\bullet+}$ (464).

Conclusion

Two new adamantyl-containing 5-aminopyrazole inhibitors were synthesized. Their effects on ubiquitin-specific protease 7 were assessed.

References

1. A. Georges, E. Marcon, J. Greenblatt, L. Frappier, Sci. Rep. 2019, 9, 15664.

2. Z. Wang, W. Kang, Y. You, J. Pang, H. Ren, Z. Suo, H. Liu, Y. Zheng, Front. Pharmacol. 2019, 10, 427.

3. H.-B. Park, S. Hwang, K.-H. Baek, Cell Death Dis. 2022, 13, 698.

4. G. Saha, S. Roy, M. Basu, M. K. Ghosh, Biochim. Biophys. Acta Rev. Cancer.2023, 1878, 188903.

5. S.-M. Qi, G. Cheng, X.-D. Cheng, Z. Xu, B. Xu, W.-D. Zhang, J.-J. Qin, Front. Cell Dev. Biol. 2020, 8, 233.

6. T. An, Y. Gong, X. Li, L. Kong, P. Ma, L. Gong, H. Zhu, C. Yu, J. Liu, H. Zhou, B. Mao, Y. Li, Biochem. Pharmacol. 2017, 131, 29.

7. A. Colleran, P. E. Collins, C. O'Carroll, A. Ahmed, X. Mao, B. McManus, P. A. Kiely, E. Burstein, R. J. Carmody, Proc. Natl. Acad. Sci. U.S.A 2013,110, 618.

8. J. van Loosdregt, V. Fleskens, J. Fu, A. B. Brenkman, C. P. J. Bekker, C. E. G. M. Pals, J. Meerding, C. R. Berkers, J. Barbi, A. Gröne, A. J. A. M. Sijts, M. M. Maurice, E. Kalkhoven, B. J. Prakken, H. Ovaa, F. Pan, D. M. W. Zaiss, P. J. Coffer, Immunity 2013, 39, 259.

9. M. S. Song, L. Salmena, A. Carracedo, A. Egia, F. Lo-Coco, J. Teruya-Feldstein, P. P. Pandolfi, Nature 2008, 455, 813.

10. G. J. Valles, I. Bezsonova, R. Woodgate, N. W. Ashton, Front. Cell Dev.Biol. 2020, 8, 717.

11. J. Lu, H. Zhao, C. Yu, Y. Kang, X. Yang, Front. Pharmacol. 2021, 12, 648491.

12. M. Ye, J. He, J. Zhang, B. Liu, X. Liu, L. Xie, M. Wei, R. Dong, K. Li, D. Ma, K. Dong, Cancer Biomark. 2021, 31, 107.

13. S. Collaud, V. Tischler, A. Atanassoff, T. Wiedl, P. Komminoth, C. Oehlschlegel, W. Weder, A. Soltermann, BMC Cancer 2015, 15, 74.

14. D. Masuya, C. Huang, D. Liu, T. Nakashima, H. Yokomise, M. Ueno, N. Nakashima, S. Sumitomo, J. Pathol. 2006, 208, 724.

15. P. Liu, Z. Chen, Y. Guo, Q. He, C. Pan, Eur. J. Med. Chem. 2025, 287, 117324.

16. P. Li, Y. Liu, H.-M. Liu, Expert Opin. Ther. Pat. 2022, 32, 753.

17. R. I. Oliveira, R. A. Guedes, J. A. R. Salvador, Front. Chem. 2022, 10, 1005727.

18. L. Nininahazwe, B. Liu, C. He, H. Zhang, Z.-S. Chen, Drug Discov. Today 2021, 26, 490.

19. B. M. Kessler, Expert Opin. Ther. Pat. 2014, 24, 597.

20. A. P. Turnbull, S. Ioannidis, W. W. Krajewski, A. Pinto-Fernandez, C. Heride, A. C. L. Martin, L. M. Tonkin, E. C. Townsend, S. M. Buker, D. R. Lancia, J. A. Caravella, A. V. Toms, T. M. Charlton, J. Lahdenranta, E. Wilker, B. C. Follows, N. J. Evans, L. Stead, C. Alli, V. V. Zarayskiy, A. C. Talbot, A. J. Buckmelter, M. Wang, C. L. McKinnon, F. Saab, J. F. McGouran, H. Century, M. Gersch, M. S. Pittman, C. G. Marshall, T. M. Raynham, M. Simcox, L. M. D. Stewart, S. B. McLoughlin, J. A. Escobedo, K. W. Bair, C. J. Dinsmore, T. R. Hammonds, S. Kim, S. Urbé, M. J. Clague, B. M. Kessler, D. Komander, Nature 2017, 550, 481.

21. O. Tavana, W. Gu, J. Mol. Cell. Biol. 2017, 9, 45.

22. I. Lamberto, X. Liu, H.-S. Seo, N. J. Schauer, R. E. Iacob, W. Hu, D. Das, T. Mikhailova, E. L. Weisberg, J. R. Engen, K. C. Anderson, D. Chauhan, S. Dhe-Paganon, S. J. Buhrlage, Cell. Chem. Biol. 2017, 24, 1490.

23. I. E. Wertz, J. M. Murray, Drug Discov. Today Technol. 2019, 31, 109.

24. O. Bruno, C. Brullo, F. Bondavalli, S. Schenone, S. Spisani, M. S. Falzarano, K. Varani, E. Barocelli, V. Ballabeni, C. Giorgio, M. Tognolini, Bioorg. Med. Chem. 2009, 17, 3379.

25. B. Marengo, E. Meta, C. Brullo, C. D. Ciucis, R. Colla, A. Speciale, O. Garbarino, O. Bruno, C. Domenicotti, Oncotarget 2023, 14, 129.

26. E. Morretta, C. Brullo, R. Belvedere, A. Petrella, A. Spallarossa, M. C. Monti, Int. J. Mol. Sci. 2023, 24, 9200.

27. E. Meta, C. Brullo, A. Sidibe, B. A. Imhof, O. Bruno, Eur. J. Med. Chem. 2017, 133, 24.

28. M. Lusardi, B. Wehrle-Haller, A. Sidibe, M. Ponassi, E. Iervasi, C. Rosano,

ECONOMIC SCIENCES

СКРИНИНГ, МАСШТАБИРОВАНИЕ И ГРАНИЦА РАСПРЕДЕЛЕНИЯ КАПИТАЛА: МОДЕЛЬ СТРУКТУРЫ ИННОВАЦИОННОГО КАПИТАЛА И ДОЛГОСРОЧНОГО РОСТА ПРОИЗВОДИТЕЛЬНОСТИ

Амиров А.
Orcid ID: 0009-0003-4167-4431
диссертант по специальности «Экономическая теория»,
Институт экономики Министерства науки и образования Азербайджанской Республики, г. Баку

SCREENING, SCALING, AND THE ALLOCATION FRONTIER: A MODEL OF INNOVATION-CAPITAL COMPOSITION AND LONG-RUN PRODUCTIVITY GROWTH

Amirov A.
Orcid ID: 0009-0003-4167-4431
PhD Candidate in Economic Theory,
Institute of Economics, Ministry of Science and Education of the Republic of Azerbaijan, Baku
DOI: 10.5281/zenodo.22107046

АННОТАЦИЯ

В литературе по финансам и инновациям связь между ними традиционно рассматривается как монотонная: увеличение капитала, направляемого в инновационные фирмы, должно повышать долгосрочный рост производительности. В статье показано, почему это не обязательно так. Разработана модель шумпетерианского роста, в которой финансисты осуществляют скрининг разнородных кандидатов на инновации перед их финансированием, сталкиваясь при этом с реальным компромиссом между издержками и скоростью: более строгий отбор повышает среднее качество финансируемых проектов, но замедляет темп, с которым капитал достигает границы инноваций. Темп сбалансированного роста является строго вогнутой, одновершинной функцией интенсивности скрининга — «границей распределения» — с технологически заданным максимумом и децентрализованным равновесием, которое всегда строго не достигает этого максимума на величину, определяемую соотношением издержек скрининга и ценности правильного отбора. Калибровка модели, сочетающая данные ОЭСР о производительности, литературу о наценках и иллюстративное целевое значение интенсивности скрининга, показывает, что это отставание обходится калиброванной экономике примерно в четыре процента от достижимого темпа роста. Сравнительная статика также показывает, что «увеличение инновационного финансирования» — не единое явление: удешевление скрининга и рост неоднородности претендентов повышают рост, приближая экономику к собственному максимуму границы, тогда как увеличение самого потока кандидатов повышает рост, отдаляя экономику от неизменного максимума.

ABSTRACT

The literature on finance and innovation treats the relationship as monotone: more capital directed at innovative firms should raise long-run productivity growth. This paper shows why that need not hold. A Schumpeterian growth model is developed in which financiers screen heterogeneous candidate innovations before funding them, subject to a genuine cost-speed tradeoff: stricter screening raises the average quality of funded projects but slows the rate at which capital reaches the innovation frontier. The balanced-growth rate is a strictly concave, single-peaked function of screening intensity — an Allocation Frontier — with a technologically determined peak and a decentralized equilibrium that always falls strictly short of it, by an amount governed by the cost of screening relative to the value of getting it right. A calibration combining OECD productivity data, the markup literature, and an illustrative target for screening intensity finds this shortfall costs the calibrated economy roughly four percent of its attainable growth rate. Comparative statics further show that “more innovation finance” is not one thing: cheaper screening and greater entrant heterogeneity both raise growth by moving the economy toward its own frontier peak, while a larger raw flow of candidates raises growth by moving it away from an unchanging one.

Ключевые слова: шумпетерианский рост, скрининг, неэффективное распределение ресурсов, инновационное финансирование, лестницы качества.

Keywords: Schumpeterian growth, screening, misallocation, innovation finance, quality ladders.

1. Formulation of the Problem

Over the past three decades, advanced economies have channelled a growing share of capital toward innovation-intensive firms — through venture capital, growth-stage private equity, corporate R&D financing, and public innovation-support programmes. Much of this expansion rests on an implicit premise, shared by a large empirical literature linking financial development to growth [1, 2], that the relationship between innovation finance and growth is monotone: more capital directed at innovation mechanically produces more innovation, and hence faster long-run total factor productivity (TFP) growth.

The empirical record is less tidy than this premise suggests. Periods of abundant innovation finance have coincided with declining apparent hit rates among

funded ventures, with capital chasing an expanding population of derivative or low-quality projects. Conversely, episodes of credit scarcity have visibly starved well-vetted, high-quality innovations of the capital needed to scale. Both patterns sit uneasily with a theory in which only the level of innovation finance matters.

This paper argues that the missing ingredient is composition, not level. The financial sector's screening intensity — the care and selectivity with which capital is matched to innovation projects — is itself a choice, and that choice carries a genuine cost-speed tradeoff. Screening harder raises the average quality of funded projects, but it also slows the rate at which capital reaches the innovation frontier, because screening consumes real resources and time that would otherwise go toward deploying capital at scale. Once this tradeoff is made explicit, the relationship between innovation-finance intensity and long-run growth need not be monotone: it can instead be hump-shaped, with an interior, productivity-maximising composition of innovation capital — the Allocation Frontier.

2. Analysis of Recent Research and Publications

This paper draws on, and departs from, three literatures: Schumpeterian growth theory, the theory and evidence on financial screening and delegated monitoring, and the literature on misallocation and aggregate productivity.

The model developed here builds on the quality-ladder framework of [3], in which aggregate growth is the product of an innovation arrival rate and the size of the quality improvement upon each successful innovation. Subsequent work has enriched this framework considerably: [4] embed it in a firm-level model of multi-product innovation; [5] survey the broader research programme; and [6] connect creative destruction explicitly to reallocation across incumbents and entrants. In this tradition, the innovation arrival rate and the size of the quality jump are typically governed by firms' own R&D or entry decisions, with the financial sector, where modelled at all, supplying funds passively at a market price. This paper's departure is to make the financial sector an active participant: financiers choose how intensively to screen innovation projects before funding them, and this choice affects both the arrival rate and the average quality jump simultaneously.

The direct predecessor for embedding finance inside a quality-ladder engine is [7], who introduce imperfect creditor protection into a multicountry Aghion-Howitt economy: countries above a critical level of financial development converge to the world growth frontier, while those below it are permanently slower — a threshold with no downside to sitting on the favourable side of it. A long empirical tradition [1, 7, 8, 9] finds that financial development supports growth; in the innovation context specifically, [2] show that venture capital funding raises patenting. More recent work complicates the picture: [10] find that beyond a point, growth in the financial sector is associated with slower productivity growth, particularly in R&D-intensive industries, while [11] document a hump-shaped relationship between financial depth and growth across countries. Both motivate non-monotonicity empirically, at the level of aggregate finance, without modelling an innovation-specific mechanism.

On the theory side, the rationale for costly screening is classical: [12] on costly state verification and [13] on delegated monitoring both establish why intermediaries screen and monitor rather than lend passively, a logic extended to capital-constrained intermediaries by [14]. The closest existing model is [15], who embed a costly-state-verification technology into a growth model in which monitoring efficacy depends on the resources devoted to it; as monitoring improves, capital is redirected from undeserving to deserving firms and the economy approaches its first-best — again monotonically, since better monitoring there raises accuracy without ever reducing the number of projects funded. In the venture capital literature, [16] document that pre-investment screening is one of the three channels VCs use to manage the entrepreneur relationship, and [17] show that "hot" venture capital markets lead investors to fund riskier, more novel startups — higher failure rates, but more patents conditional on success — because abundant capital lowers the cost of experimenting on less-proven ideas rather than simply lowering the funding bar.

The paper's second building block is the misallocation-accounting framework of [18] and [19], in which dispersion in the marginal product of capital across producers translates directly into an aggregate TFP shortfall. [20] tie this dispersion to financial frictions, and [21] show that imperfect information about firm-level productivity — a close cousin of imperfect screening — can generate the same kind of misallocation. [22] connects markup dispersion and misallocation to growth dynamically within a Schumpeterian setting, methodologically the closest antecedent to the approach taken here.

3. Identification of Previously Unresolved Parts of the General Problem

What this literature does not do, across all three strands, is embed screening intensity as an endogenous choice inside a general-equilibrium growth model with an explicit cost-speed tradeoff, or derive a growth-maximising screening intensity. In each misallocation study cited above, the degree of misallocation is a consequence of an exogenous friction, not an equilibrium object; in the finance-and-screening literature, monitoring or institutional quality is either exogenous or, when it improves, unambiguously good, with no framework admitting the possibility that screening could become excessive.

This last point is worth dwelling on, since a companion analysis by the present author [23] argues, informally and using a Credit × Institutional Quality interaction, that credit expansion raises growth where institutional quality and screening sophistication are already strong, and can produce an innovation trap where they are weak — concluding that financial architecture should prioritise meritocratic selection of ideas over the preservation of existing firm structures. That analysis anticipates the present paper's starting point, but its structure differs in three respects. First, institutional quality and screening sophistication are treated as exogenous context, not as a choice made by the financial sector. Second, the relationship is monotonic:

more institutional quality is always weakly better, so the framework has no analogue of screening being excessive. Third, the resulting nonlinearity is a one-sided threshold — a trap below some level, with no corresponding penalty above it — rather than an interior optimum. The literature, including that companion analysis, therefore still lacks a model in which screening intensity is chosen by the financial sector itself, embedded in a general-equilibrium growth framework, capable of generating and characterising an interior optimum rather than simply asserting one.

4. Purpose of the Article

The purpose of this article is to develop a tractable Schumpeterian growth model in which financiers' screening intensity is an endogenous choice subject to a genuine cost-speed tradeoff, to characterise analytically when an increase in innovation-oriented finance raises long-run productivity growth and when it lowers it, and to quantify the resulting Allocation Frontier and its comparative statics using a calibration disciplined by published aggregates. Specifically, the paper: (i) extends the quality-ladder framework of [3] to allow heterogeneous entrant quality and an explicit financier screening technology; (ii) characterises the balanced growth path and proves that the growth-maximising screening intensity is technologically determined while the equilibrium intensity always falls strictly short of it; (iii) connects the model to the misallocation-accounting logic of [18]; and (iv) derives and quantifies comparative statics with respect to screening cost, entrant-quality dispersion, and credit supply.

5. Presentation of the Main Material

This section lays out the model, its equilibrium, its central proposition, and its quantitative implementation. Notation: γ denotes the (random) quality step of a candidate innovation, with support $[\gamma_{min}, \gamma_{max}]$; Λ denotes the exogenous raw arrival rate of candidates, to keep it visually distinct from $\lambda(s)$, the endogenous rate of funded innovations.

5.1. Model Environment

This subsection lays out the economic environment, following a standard Schumpeterian quality-ladder structure in the tradition of [3]. To fix notation and establish a reference growth rate, the innovation arrival rate and quality step are here held fixed at exogenous values, λ and γ; Section 5.2 relaxes both.

Households. A representative household of measure one has lifetime utility

$$U_0 = \int_0^\infty e^{-\rho t} \cdot \frac{C_t^{1-\theta} - 1}{1-\theta} \, dt, \rho > 0, \theta > 0 \quad (1)$$

where C_t is consumption, ρ is the discount rate, and θ is the coefficient of relative risk aversion (the inverse of the intertemporal elasticity of substitution). The household holds a diversified portfolio of claims on the economy's monopolist producers, earning the market return r_t. Standard optimization gives the Euler equation

$$\frac{\frac{dC_t}{dt}}{C_t} = \frac{1}{\theta}(r_t - \rho) \quad (2)$$

There is no separate physical capital stock in this economy, and no labor margin: growth is driven entirely by quality upgrading, and "capital" throughout refers to the financial resources directed at funding innovation projects, not to a Solow-style accumulable factor. This is a deliberate simplification, standard in the baseline Schumpeterian tradition, that keeps the analysis focused on the allocation margin that is the paper's object of interest.

Final good production. A continuum of intermediate varieties $j \in [0,1]$ is combined competitively into a single final good,

$$Y_t = \int_0^1 A_{jt}^{1-\alpha} x_{jt}^{\alpha} \, dj, \alpha \in (0,1) \quad (3)$$

where A_{jt} is the current productivity ("quality rung") of variety j and x_{jt} is the quantity used. Profit maximization by the competitive final-good sector, taking the final good as numeraire, yields inverse demand for each variety,

$$p_{jt} = \alpha A_{jt}^{1-\alpha} x_{jt}^{\alpha-1} \quad (4)$$

Intermediate goods and creative destruction. Each variety j is supplied by a single incumbent monopolist — the current quality leader — who produces one-for-one from the final good at unit marginal cost. Facing demand (4), the incumbent chooses x_{jt} to maximize $p_{jt} x_{jt} - x_{jt}$, giving

$$x_{jt}^* = \kappa A_{jt}, \kappa \equiv (\alpha^2)^{\frac{1}{1-\alpha}} \quad (5)$$

a constant equilibrium markup price $p^* = 1/\alpha$, and profit linear in quality,

$$\pi_{jt} = \pi A_{jt}, \pi \equiv \frac{1-\alpha}{\alpha} \kappa \quad (6)$$

Substituting (5) into (3), aggregate output is linear in the economy's aggregate quality index,

$$Y_t = \kappa^{\alpha} A_t, A_t \equiv \int_0^1 A_{jt} \, dj \quad (7)$$

Quality evolves through creative destruction. In each sector, entrants arrive and, if funded, displace the incumbent and raise the quality rung multiplicatively: $A_{jt} \to \gamma A_{jt}$, $\gamma > 1$. Held fixed for now, this event occurs at Poisson rate λ per sector, independently across the continuum. By a standard law-of-large-numbers argument, the aggregate quality index then evolves deterministically,

$$g \equiv \frac{\frac{dA_t}{dt}}{A_t} = \lambda(\gamma - 1) \quad (8)$$

which is also the balanced-growth rate of output, consumption, and profit, all of which are linear in A_t. Equation (8) is the growth rate this economy would deliver under the canonical Aghion-Howitt assumption that the arrival rate and step size are exogenous. It is the reference point against which the rest of this section measures the effect of making both objects endogenous to a screening choice.

Asset values and the price of innovation. Because A_{jt} is constant between innovations, the value of incumbency in sector j, $V(A_{jt})$, satisfies the flow no-arbitrage condition

$$r \cdot V(A_{jt}) = \pi_{jt} - \lambda \cdot V(A_{jt}) \quad (9)$$

reflecting the risk of displacement at rate λ, which destroys the incumbent's claim. Using (6) and guessing

(and verifying) a solution linear in quality, V(A_jt) = v·A_jt,

$$v = \frac{\pi}{r + \lambda} \quad (10)$$

The object vγA_jt — the value of becoming the new incumbent in a sector currently at quality A_jt — is the prize an entrant (or the financier funding it) receives upon a successful innovation, and is the key input into the funding decision formalized in Section 5.3.

Resource constraint and the reference interest rate. Final output is allocated between consumption and intermediate-input production:

$$Y_t = C_t + \int_0^1 x_{jt}\, dj = C_t + \kappa A_t \quad (11)$$

Since every term in (11) is linear in A_t, consumption grows at the same balanced rate g as output along a balanced growth path, and the Euler equation (2) then pins down the reference interest rate,

$$r = \rho + \theta g \quad (12)$$

The rest of this section introduces entrant heterogeneity and an explicit financier screening technology, which jointly determine λ and γ — and, through them, g, r, and v — as equilibrium objects rather than free parameters. A resource cost of screening is added to (11). The connection to the misallocation-accounting framework of [18], developed in Section 5.4, runs not through a distortion internal to (5) — funded projects turn out to be efficiently sized given their own realized quality — but through the dispersion of realized quality among funded projects, which is exactly what a Hsieh-Klenow-style exercise applied to newly funded sectors would recover, and which the screening margin governs directly.

5.2. Entrant Heterogeneity and the Screening Technology

Section 5.1 treated the quality step γ as a single exogenous parameter, common to every innovation. This subsection relaxes that assumption. Candidates for funding differ in quality, and what an incumbent's displacement is actually worth to the economy depends on which candidates get funded — which is where the financial sector's screening choice enters.

Entrant heterogeneity. In each sector, candidates arrive according to a Poisson process with rate Λ, an exogenous flow of raw innovation opportunities; only the financing margin, not the arrival of ideas itself, is endogenous in this paper. Each candidate carries a quality step γ, drawn independently from a uniform distribution on $[\gamma_{min}, \gamma_{max}]$, with

$$\gamma_{min} < 1 < \gamma_{max} \quad (13)$$

The restriction $\gamma_{min} < 1$ is the substantive assumption driving the paper's central result. A candidate with γ<1 would, if funded, displace a working incumbent with something worse: a mismanaged venture, a product without a market, outright failure. This is not a technical nicety; it is the empirically ordinary case that most funded ventures do not succeed. Absent such candidates — if every funded project were guaranteed to be a weak improvement at worst — a planner would always want to fund every candidate as fast as possible, and screening would be pure waste with no offsetting benefit. It is the presence of genuinely value-destroying candidates in the pool that makes screening valuable rather than merely costly.

The screening technology. Because sectors are symmetric, it suffices to characterize the common screening intensity chosen in a representative sector; whether this is implemented by a single economy-wide financial sector or by competitive, identical financiers in each sector does not matter for what follows. A financier chooses a screening intensity s∈[0,1], applied uniformly to arriving candidates. Screening intensity maps into a funding threshold Γ(s),

$$\Gamma(s) = \gamma_{min} + s\Delta, \Delta \equiv \gamma_{max} - \gamma_{min} \quad (14)$$

so that a candidate is funded if and only if γ≥Γ(s); s=0 funds every candidate, and s=1 funds none. Because γ is uniformly distributed, s has an immediate interpretation as the fraction of candidates the threshold rejects: F(Γ(s))=s.

Reliably enforcing a stricter threshold is costlier: distinguishing a candidate just above a demanding bar from one just below it requires more careful evaluation than screening out an obviously hopeless one. Evaluating every arriving candidate at intensity s costs

$$\chi(s) = \frac{\zeta}{2} s^2 A_{jt}, \zeta > 0 \quad (15)$$

units of the final good, paid regardless of whether the candidate is ultimately funded. Cost scales with the current quality of the sector under evaluation: vetting a venture that could displace a larger incumbent is proportionately more involved, and this scaling is also what keeps the screening choice from drifting as the sector — and the economy — grows, so that it can be a genuine constant along a balanced growth path. The quadratic form is adopted for tractability; ζ indexes the cost of screening and is the paper's primary comparative-static object below.

Three mechanical consequences of (14) follow directly and are used repeatedly below. First, the effective arrival rate of funded innovations,

$$\lambda(s) = \Lambda[1 - F(\Gamma(s))] = \Lambda(1 - s) \quad (16)$$

falls linearly in screening intensity: this is the throughput cost of screening — capital reaches the frontier more slowly the choosier the financial sector becomes. Second, the average quality step among funded candidates,

$$\gamma(s) = E[\gamma \mid \gamma \geq \Gamma(s)] = \\ = \frac{\gamma_{min} + \gamma_{max}}{2} + \frac{s\Delta}{2} \quad (17)$$

rises linearly in screening intensity: this is the selection benefit. Third, the dispersion of realized quality among funded candidates,

$$Var(\gamma \mid s) = \frac{\Delta^2 (1 - s)^2}{12} \quad (18)$$

falls in screening intensity — a stricter bar not only raises the average outcome but also compresses it, which is the channel that connects this model to the misallocation-accounting logic of [18] below.

Total screening expenditure per sector per unit time is Λ·χ(s), a resource cost added to (11). Equations (16) and (17) already point toward the central tension: raising s trades a higher average payoff per funded innovation against fewer innovations being funded at all, on top of the direct resource cost (15). Section 5.3 shows how a financier weighs this tradeoff; Section 5.4

uses the resulting choice, together with (18), to characterize the balanced growth path and state this paper's central result.

5.3. The Financier's Problem and Equilibrium Screening Intensity

Substituting (16) and (17) into the growth definition from Section 5.1, the balanced-growth rate that a screening intensity s would deliver, if entrenched permanently, is

$$g(s) = \lambda(s)[\gamma(s) - 1] = \Lambda(1-s)\left[m + \frac{s\Delta}{2}\right], \quad m \equiv \frac{\gamma_{min} + \gamma_{max}}{2} - 1 \tag{19}$$

where m is the average net quality improvement in the entrant pool. Maintained throughout is m>0: the average candidate is quality-improving, even though, by (13), a positive fraction of individual candidates are not. Expanding (19) shows g(s) is a strictly concave quadratic in s, maximized at

$$s^{FB} = \frac{1}{2} - \frac{m}{\Delta} = \frac{1 - \gamma_{min}}{\Delta} \tag{20}$$

the value of s at which the threshold (14) exactly equals 1 — screening out precisely the value-destroying candidates and no others. This is the growth-maximizing screening intensity absent any cost of screening, and the natural first-best benchmark against which the financier's actual choice is compared.

A financier, price-taking in the value ratio v (consistency of v with the aggregate outcome is verified in Section 5.4), chooses s to maximize the expected net value flow from screening a representative sector. Using the cost function (15), the financier's problem is

$$max_{s\in[0,1]}\, \Pi(s) = A_{jt}\left[v \cdot g(s) - \Lambda \cdot \frac{\zeta}{2} s^2\right] \tag{21}$$

A_jt scales the whole objective and drops out of the optimality condition, so the chosen intensity is common across sectors regardless of their individual histories — a necessary condition for a balanced growth path in which s itself is constant. The first-order condition, using (19)'s expansion g′(s) = Λ[(Δ/2 − m) − sΔ], gives

$$s^*(v) = s^{FB} \cdot \frac{v\Delta}{v\Delta + \zeta} \tag{22}$$

which is strictly increasing in v, strictly decreasing in ζ, and satisfies 0 < s*(v) < s^FB < 1 for every finite ζ>0 and v>0 (Appendix A verifies the second-order condition and the box constraint). Equilibrium screening is always interior to [0, s^FB] under this cost structure: financiers never fully close the gap to the cost-free benchmark, and the size of the shortfall, s^FB − s*(v), is governed entirely by the ratio ζ/(vΔ) — the cost of screening relative to the stakes of getting it right.

Two limits are worth stating explicitly. As ζ→0, s*(v)→s^FB: costless verification lets financiers implement the first-best threshold exactly. As ζ→∞, s*(v)→0: prohibitively costly verification collapses the financial sector back to funding everything indiscriminately, the opposite extreme. Equation (22) is the financier's best response, expressed as a function of the value ratio v; Section 5.4 closes the model by determining v itself along the balanced growth path and states the paper's central proposition.

5.4. Balanced Growth Path: The Allocation Frontier

Closing the loop. Sections 5.1 and 5.3 leave one piece of circularity unresolved: the financier's best response (22) takes the value ratio v as given, while v itself — via (10) and (12) — depends on the balanced-growth rate that the financier's own choice of s determines. Combining (10) and (12),

$$v(s) = \frac{\pi}{\rho + \theta \cdot g(s) + \lambda(s)} \tag{23}$$

which is continuous and strictly positive for every s∈[0,1], since g(s)≥0 and λ(s)≥0 throughout. A balanced growth path is a screening intensity s* satisfying

$$s^* = s^{FB} \cdot \frac{v(s*)\Delta}{v(s*)\Delta + \zeta} \tag{24}$$

Proposition 1 (Existence). A balanced growth path exists: equation (24) has a solution s*∈(0, s^FB).

Proof sketch. The right-hand side of (24), as a function of s, is continuous on [0,1] (a composition of continuous functions) and takes values strictly inside (0, s^FB)⊂(0,1) for every s, by the bound established in Section 5.3. Evaluating the difference between the right-hand side and s at the endpoints: at s=0 the right-hand side is strictly positive, so the difference is positive; at s=1 the right-hand side is strictly below s^FB<1, so the difference is negative. By the intermediate value theorem, the difference vanishes somewhere in (0,1), and any such zero automatically lies in (0, s^FB). Full details are in Appendix A.

Uniqueness is not guaranteed in general, since g(s) is concave but v(s) need not be monotonic over the full range s∈[0,1]; a sufficient condition for a unique balanced growth path — bounding the discount-rate sensitivity θ relative to the curvature of g — is given in Appendix A, and the calibrated economy in Section 5.6 has a unique, numerically verified balanced growth path throughout.

The Allocation Frontier. The object of central interest is the mapping from screening intensity to the balanced-growth rate, g(·), given in closed form by (19).

Proposition 2 (The Allocation Frontier). The balanced-growth rate g(s) is a strictly concave, single-peaked function of screening intensity on [0,1], maximized at the interior point s^FB = $(1-\gamma_{min})/\Delta \in (0,1)$. For any finite screening cost ζ>0, the balanced-growth-path intensity s* from Proposition 1 satisfies 0<s*<s^FB, and delivers a growth rate strictly below the maximum attainable, g(s*)<g(s^FB), with the shortfall vanishing as ζ→0.

Proof. Concavity and the location of the peak follow from the quadratic expansion of (19), used already to derive (20). That s*<s^FB follows immediately from (22): for any finite ζ>0 and v>0, the multiplying factor vΔ/(vΔ+ζ) is strictly less than one, so any fixed point of (24) inherits the same strict inequality. Since g is strictly concave with a unique maximizer at s^FB, any s≠s^FB satisfies g(s)<g(s^FB), and continuity of s*(·) in ζ (Appendix A) gives s*→s^FB, hence g(s*)→g(s^FB), as ζ→0.

This is the paper's central result. Innovation-finance intensity — indexed here by screening intensity s — does not raise long-run TFP growth monotonically: g(s) rises while screening is removing genuinely value-

destroying candidates from the funded pool, and falls once it starts rejecting candidates that, while modest, were still worth funding. The peak of this hump, s^FB, is a purely technological object, pinned down by the lower tail of entrant quality alone; it does not depend on the cost of screening, the value of innovation, or the discount rate. The equilibrium position on the frontier, s*, does depend on all of these, through (22) and (24), and sits strictly to the left of the peak for any positive screening cost — the financial sector, in this economy, always screens somewhat less than the growth-maximizing amount, because doing more would cost more than the marginal selectivity is worth.

Nothing here reflects a market failure: the financier's objective (21) already internalizes the full social value of the marginal quality unit, so the gap s^FB−s* is the efficient consequence of a real resource cost, not an externality this paper's baseline is equipped to detect. A fuller welfare treatment allowing for the classic appropriability and business-stealing externalities of the Schumpeterian tradition is a natural extension this paper does not pursue.

Dispersion and the connection to Hsieh-Klenow. Section 5.1 flagged that imperfect screening would connect this model to the misallocation-accounting logic of [18]. The connection runs through (18), not through a distortion internal to (5): because screening reveals a candidate's true quality before funding, every project that is financed is efficiently sized given its own realized γ, and nothing in this baseline corresponds to a genuine capital-sizing wedge. What screening intensity governs instead is the cross-sectional spread of realized outcomes among funded projects — precisely the object a Hsieh-Klenow-style exercise, applied to a cross-section of newly funded sectors, would recover as dispersion in realized returns.

The caution this implies is worth stating plainly: in this economy, a decline in measured dispersion as screening tightens reflects a more selective, lower-throughput funding process, not a reduction in any genuine misallocation, since funded projects are always efficiently sized here. Cross-sectional dispersion in realized returns is not, by itself, evidence of misallocation when the population being sampled is itself heterogeneous by design — a caution that generalizes beyond this specific model, and one directly relevant to any empirical exercise that reads screening intensity off of realized dispersion in funded-project outcomes.

5.5. Comparative Statics

This subsection signs the response of equilibrium screening intensity and growth to three primitives: the cost of screening, the dispersion of entrant quality, and the supply of credit. Some results are fully closed-form; others require a mild regularity condition or are sharpened numerically in Section 5.7.

Screening cost. Holding v fixed, (22) already shows $\partial s^*/\partial\zeta<0$: costlier screening induces less of it. In the full balanced-growth-path equilibrium, the same conclusion follows from the implicit function theorem applied to (24), provided the equilibrium is locally regular — that the right-hand side of (24) crosses the 45-degree line from above, the generic case whenever a crossing exists, and verified numerically for the calibrated economy below. Under this condition, $ds^*/d\zeta<0$. Since Proposition 2 places s* strictly to the left of the peak s^FB, and g is strictly increasing on [0, s^FB), a fall in s* moves the economy further from the peak in the direction that lowers growth: $dg(s^*)/d\zeta<0$. More expensive screening technology unambiguously lowers both equilibrium selectivity and equilibrium growth — the plainest of the paper's comparative statics, and the one closest to the intuition of the underlying cost-speed tradeoff.

Entrant dispersion. Isolating dispersion from average quality requires holding the mean of the entrant-quality distribution fixed while varying its spread. Write $\gamma_{min} = (1+m) - \Delta/2$ and $\gamma_{max} = (1+m) + \Delta/2$ for a fixed mean 1+m, so that m is held constant as Δ varies; the maintained assumption $\gamma_{min}<1$ then requires $\Delta>2m$.

Under this parameterization, $s^{FB} = 1/2 - m/\Delta$ is strictly increasing in Δ: a more heterogeneous entrant pool, at a fixed average quality, calls for rejecting a larger fraction of candidates, because the extra mass added by a mean-preserving spread sits partly below the existing floor. The maximum attainable growth rate itself,

$$g(s^{FB}) = \Lambda\left[\frac{m}{2} + \frac{\Delta}{8} + \frac{m^2}{2\Delta}\right] \tag{25}$$

has derivative $dg(s^{FB})/d\Delta = \Lambda[1/8 - m^2/(2\Delta^2)]$, strictly positive throughout the admissible region $\Delta>2m$ (equalling zero only at the excluded boundary $\Delta=2m$, exactly where $\gamma_{min}=1$). Greater entrant heterogeneity, holding average quality fixed, always raises the growth ceiling on the admissible domain: a wider pool has a fatter bad tail, but also a fatter good one, and correctly screening out the former more than compensates.

The partial-equilibrium best response (22) inherits the same sign: s^FB and the multiplying factor $v\Delta/(v\Delta+\zeta)$ both rise in Δ, so s* rises too, for fixed v. The full general-equilibrium response of g(s*) — combining a rising ceiling with a shifting gap s^FB−s* — is not signed in closed form here; Section 5.7 shows it is positive throughout the calibrated region.

Credit supply. The natural credit-supply parameter in this model is Λ, the raw arrival rate of candidates reaching the financing stage. A rise in Λ raises g(s) and λ(s) proportionally at any fixed s, and through (23) this raises the denominator ρ+θg(s)+λ(s) — so v(s) falls as Λ rises, for any fixed s. More candidates competing for the same quality ladders shortens the expected tenure of any given incumbent and raises the discount rate, both of which make winning less valuable in present-value terms.

Because s*(v) is increasing in v (Section 5.3), a fall in v pulls equilibrium screening down: an expansion of credit supply, by lowering the capitalized value of a successful innovation, induces financiers to screen less intensively, not more. This works against the direct, mechanical scale effect of more candidates arriving in the first place, and the two forces have to be weighed against each other numerically; Section 5.7 shows the scale effect dominates for the calibrated parameter region, so that credit-supply expansions raise growth on net, but by less than a model without this value channel would predict.

Section 5.6 turns to calibration, disciplining Λ, γ_{min}, γ_{max}, ζ and the household parameters ρ, θ against

OECD aggregates and published elasticities, so that the comparative statics above can be evaluated at economically meaningful parameter values rather than left as qualitative signs.

5.6. Calibration

This subsection disciplines the model's seven parameters against a mix of standard macroeconomic convention, published aggregates, and — where no clean external counterpart exists — transparent, internally consistent illustrative targets.

Standard conventions and external targets. Three parameters require no extended discussion. The discount rate is set to $\rho=0.03$ and relative risk aversion to $\theta=2$, both standard in quantitative macroeconomics. The markup parameter is set to match an aggregate markup of 20% over marginal cost, $1/\alpha=1.20$ — a moderate value given how contested this moment is in the literature, which ranges from [24]'s widely cited estimates of aggregate U.S. markups rising toward roughly two-thirds above marginal cost in the most recent period, to considerably more conservative estimates near 10% once the full range of operating costs, not just cost of goods sold, is treated as variable. This gives $\alpha\approx0.833$, and $\kappa\approx0.112$, $\pi\approx0.022$ from (5)-(6).

The balanced-growth rate is targeted at $g^*=1.0\%$ per year, consistent with total factor productivity growth in the United States and other OECD economies over recent decades, which has generally registered between roughly 0.4% and 1.1% depending on the period, with the 2010s toward the lower end of that range and the 1990s-2000s toward the upper end.

The entrant-quality and screening-cost parameters. The entrant-quality distribution and the screening-cost parameter have no direct empirical counterpart in existing data and are calibrated illustratively. Two considerations discipline the choice. First, the model requires $s^{FB}<1/2$ whenever the average candidate is quality-improving (Section 5.3), so any calibration must respect this bound. This rules out directly importing the extreme selectivity documented for formal venture capital, where top-tier funds reportedly finance well under 1% of the pitches they review — a useful illustration of how selective the narrowest slice of the innovation-finance funnel can be, but not a plausible target for the much broader population this model's screening margin represents in aggregate: business-loan applications, corporate R&D proposals, and angel and seed financing, alongside formal VC. Second, subject to that bound, the average net quality improvement in the entrant pool, m, and the first-best rejection fraction, s^{FB}, are set to values broadly consistent with the range of step sizes used in calibrated quality-ladder models: $m=0.10$ and $s^{FB}=0.40$.

These two choices pin down $\Delta = m/(0.5-s^{FB}) = 1.00$ via (20), and hence $\gamma_{min}=0.60$, $\gamma_{max}=1.60$: candidates range from a severe 40% quality decline to a strong 60% improvement. The remaining two parameters, Λ and ζ, are solved jointly and numerically to match two targets: the growth rate $g^*=0.01$, and a screening-intensity target of $s^*=0.28$, placing the balanced-growth-path financier 70% of the way from no screening to the first-best threshold — an illustrative middle position consistent with the paper's claim that financiers under-screen relative to s^{FB} without collapsing to either extreme.

Table 1.

Calibrated Parameters

Parameter	Value	Target / Source
ρ	0.030	Standard convention
θ	2.00	Standard convention
α	0.833	Markup $1/\alpha=1.20$ [24]
m	0.10	Illustrative; quality-ladder step-size range
s^{fB}	0.40	Illustrative; respects the $s^{fB}<0.5$ bound
γ_{min}, γ_{max}	0.60, 1.60	Implied by m, s^{fB} via (14), (20)
Λ	0.0579	Jointly solved: $g^*=0.01$, $s^*=0.28$
ζ	0.1049	Jointly solved: $g^*=0.01$, $s^*=0.28$

Table 2.

Implied Balanced Growth Path

Object	Value
s*	0.280
g*	1.00%
λ* = λ(s*)	4.17%
γ(s*)	1.240
sd(γ | s*), from (18)	0.208
v*	0.245
r*	5.00%
$g(s^{fB})$ (ceiling)	1.042%
Growth shortfall, $g(s^{fB})-g^*$	0.042 pp

Discussion. At this calibration, the balanced-growth-path financial sector rejects 28% of candidates, funds the rest at an effective rate of about 4.2% per sector per year, and captures an average quality improvement of 24% among the projects it finances. It falls short of the growth-maximizing frontier by about four hundredths of a percentage point of annual growth — roughly 4% of the maximum attainable rate — the calibrated cost of screening being a real, resource-consuming activity rather than a free filter. Section 5.7 recomputes this balanced growth path across a grid of ζ, Δ (holding m fixed), and Λ to trace out the Allocation Frontier and the comparative statics above numerically.

5.7. Quantitative Results

This subsection evaluates the calibrated model. Figure 1 plots the balanced-growth rate g(s) from (19) over the full range s∈[0,1] at the calibrated Λ, together with the three comparative statics of Section 5.5, all computed by numerically solving the full fixed point (24) rather than relying on the partial-equilibrium approximation (22).

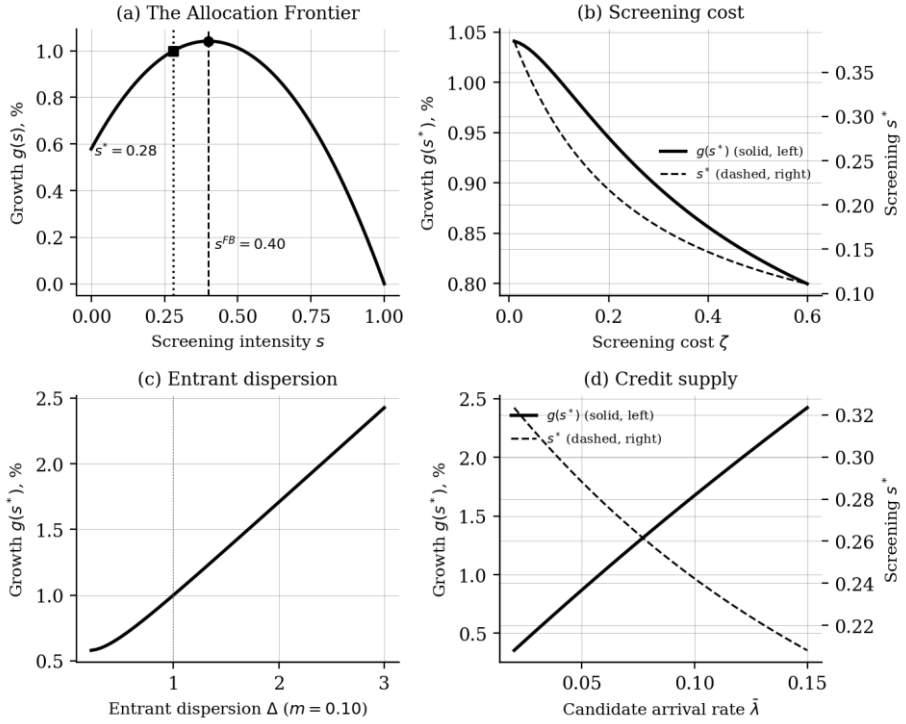


Figure 1. The calibrated Allocation Frontier and comparative statics.

The Allocation Frontier, calibrated. Panel (a) of Figure 1 shows the shape asserted in Proposition 2. Growth is 0.58% with no screening at all (s=0, when the value-destroying share of the entrant pool is funded along with everyone else), rises to a peak of 1.04% at the technological optimum s^FB=0.40, and falls back to exactly zero as s→1, when the threshold becomes so demanding that almost nothing is funded. The calibrated equilibrium, s*=0.28, sits on the rising portion of the curve, short of the peak by the 0.042-percentage-point gap already reported in Table 2.

Comparative statics, quantified. Table 3 confirms the monotone decline established analytically for screening cost. As screening becomes twenty times more expensive (from ζ=0.02 to ζ=0.40), equilibrium

selectivity falls from 37% to 15% of candidates rejected, and growth falls from 1.04% to 0.86% — a decline of eighteen basis points, close to a fifth of the baseline growth rate, from this channel alone.

Table 3.

Screening Cost

ζ	s*	g*
0.020	0.371	1.039%
0.050	0.334	1.029%
0.105 (baseline)	0.280	1.000%
0.200	0.217	0.945%
0.400	0.147	0.856%

Table 4 resolves the general-equilibrium ambiguity left open for entrant dispersion: growth rises monotonically with dispersion throughout the admissible range, more than tripling as the pool widens from just above the minimum compatible with $\gamma_{min}<1$ to a considerably wider spread. Equilibrium selectivity rises alongside it, and rises faster than the technological optimum does: s* climbs from 35% to 83% of s^FB across the range, even as s^FB itself is shifting.

Table 4.

Entrant Dispersion (m=0.10 fixed)

Λ	s^{fB}	s*	g*	Ceiling $g(s^{fB})$
0.25	0.100	0.035	0.583%	0.586%
0.50	0.300	0.161	0.681%	0.709%
1.00 (baseline)	0.400	0.280	1.000%	1.042%
1.50	0.433	0.334	1.351%	1.394%
2.50	0.460	0.383	2.067%	2.109%

Table 5 settles the tension flagged for credit supply. As the raw candidate arrival rate rises from 0.02 to 0.15 — roughly a sevenfold expansion of credit supply — equilibrium selectivity falls steadily, from 32% to 21% of candidates rejected, exactly as the value-channel argument predicts. But growth rises nearly sevenfold over the same range, from 0.35% to 2.42%: the direct scale effect of more candidates dominates the induced fall in screening by a wide margin at every point on this grid. The value channel is real and correctly signed; it is simply second-order relative to the mechanical effect of more capital reaching the financing stage in the first place, at least for the calibrated cost and dispersion parameters.

Table 5.

Credit Supply

Λ	s*	g*
0.020	0.323	0.354%
0.040	0.299	0.700%
0.058 (baseline)	0.280	1.000%
0.100	0.242	1.676%
0.150	0.208	2.424%

Taken together, these results qualify the paper's headline claim in an instructive way. Cheaper screening and wider entrant dispersion both raise growth by moving the economy toward its own frontier peak — for dispersion, s* rises from 35% to 83% of s^FB across the range in Table 4, even as s^FB itself shifts outward. Credit-supply expansion is different: it raises growth throughout, but it does so by moving the equilibrium further from an unchanging technological optimum, since s^FB does not depend on Λ at all. More innovation finance is not one thing in this model; three routes to "more" have three different relationships to the frontier they are supposedly approaching.

6. Conclusions and Proposals

This paper asked when an increase in innovation-oriented finance raises long-run productivity growth, and when it lowers it. The answer, in the Schumpeterian growth model developed here, is that the level of innovation finance is not the right object to ask the question about at all. What matters is the composition of that finance — specifically, the intensity with which the financial sector screens candidate innovations before funding them — and this composition has an interior,

productivity-maximizing optimum rather than a monotone relationship to growth.

The mechanism is a genuine cost-speed tradeoff, not merely an assumption. Because a positive share of candidate innovations would destroy value if funded, screening has a real benefit: it keeps the worst of them out of the funded pool. But because reliably enforcing a stricter bar is costlier, screening also throttles the rate at which capital reaches the innovation frontier. The balanced-growth rate that results, g(s), is a strictly concave, single-peaked function of screening intensity — the Allocation Frontier — and the decentralized financial sector, weighing the same tradeoff privately, always settles strictly short of the technologically feasible peak, by an amount that depends only on the cost of screening relative to the value of getting it right.

Three results seem worth carrying forward beyond this paper. First, the interior optimum is not a special case: Proposition 1 shows a balanced growth path with strictly interior screening always exists, for any finite screening cost. Second, the wedge between equilibrium and first-best screening in this baseline is not a market failure — it is what a real resource cost looks like once financiers already internalize the full social value of the marginal innovation — which sharpens rather than blurs the case for policy attention to screening costs specifically, as opposed to a general presumption that more oversight of financial intermediation is warranted. Third, and most useful for an applied reader, "more innovation finance" is not one thing: cheaper screening and a wider entrant pool both raise growth by moving the economy toward its own frontier peak, while a larger raw flow of candidates raises growth while moving the economy away from an unchanging one. A policy or a credit-market development that raises the volume of innovation finance without touching its composition is not obviously the same policy as one that raises its quality, even when both show up as faster growth in the aggregate data.

The model is deliberately spare, and several simplifications are worth flagging rather than hiding. The single-peaked shape of the Allocation Frontier rests on two functional-form choices adopted for tractability — a uniform entrant-quality distribution and a quadratic screening-cost function — but the underlying logic does not: any continuous entrant-quality distribution with mass on both sides of the break-even quality, combined with any strictly convex screening-cost function, generates the same tension between a selection benefit and a throughput cost. The closed-form characterization in Propositions 1-2 is specific to this paper's functional forms; the interior optimum itself is not. There is no physical capital stock or labor market, so all of the paper's content runs through the allocation of financial resources to innovation, not through a broader factor-accumulation story. Screening reveals a candidate's quality perfectly once undertaken, which is what let this paper connect the model to Hsieh-Klenow-style dispersion honestly rather than through an invented capital-sizing wedge — but a version with a genuinely noisy signal, where financiers size their investment under uncertainty about quality, would reintroduce that wedge and is a natural next step. The baseline also treats sectors symmetrically; a natural extension would allow sector-specific screening costs and entrant-quality dispersion, generating a sector-specific Allocation Frontier that could be used to study cross-sector capital reallocation directly. And the welfare analysis here stops short of the classic appropriability and business-stealing externalities that pervade the Schumpeterian tradition; nothing in this paper rules out that a fuller treatment would open a second wedge between the decentralized and planner's optima, on top of the cost-driven one characterized in Proposition 2.

An empirical counterpart for the screening-cost parameter ζ is worth flagging as future work rather than assumed away: proxies such as venture-capital due-diligence time, loan-officer effort per application, or realized rejection rates could in principle be used to locate an economy's current position on the Allocation Frontier, giving the comparative statics of Section 5.5 direct empirical content.

None of these are objections to the paper's central claim, which does not depend on any of them: as long as some share of candidate innovations would destroy value if funded, and as long as screening them out takes real time or resources, innovation finance has a productivity-maximizing composition, and treating its effect on growth as monotone will misread the data whenever an economy sits on the wrong side of it.

Appendix A. Proofs

This appendix collects the four results the main text deferred: the verification behind (22), the full proof of Proposition 1, a sufficient condition for uniqueness of the balanced growth path, and the ζ→0 limit used in the proof of Proposition 2.

A.1. Verification for Equation (22)

Lemma A.1. The financier's problem (21) is strictly concave in s on [0,1], and its unconstrained maximizer, given by (22), lies strictly inside (0, s^FB)⊂(0,1) for every finite ζ>0 and v>0; the box constraint s∈[0,1] therefore never binds.

Proof. From (19), $g(s) = \Lambda[m + s(\Delta/2 - m) - s^2\Delta/2]$, so $g''(s) = -\Lambda\Delta$ for all s. Differentiating (21) twice,

$$\Pi''(s) = A_{jt}[v \cdot g''(s) - \Lambda\zeta] = \\ = -A_{jt} \cdot \Lambda(v\Delta + \zeta) < 0$$

since A_jt, Λ, v, Δ, ζ are all strictly positive. Π is therefore strictly concave on all of $\mathbb{R}$, and its unique unconstrained maximizer solves the first-order condition, given in closed form by (22). Since $0<\gamma_{min}<1<\gamma_{max}$ by (13), $s^{FB} = (1-\gamma_{min})/\Delta \in (0,1)$; and since $v\Delta/(v\Delta+\zeta) \in (0,1)$ strictly for any v,ζ>0, (22) gives $s^*(v) \in (0, s^{FB}) \subset (0,1)$. The unconstrained maximizer already lies in the feasible set, so it solves the constrained problem too.

A.2. Proof of Proposition 1 (Existence)

Proof. Define $D(s) \equiv \rho + \theta g(s) + \lambda(s)$, $v(s) \equiv \pi/D(s)$, and $\Psi(s) \equiv s^{FB} \cdot v(s)\Delta/[v(s)\Delta+\zeta]$, so (24) reads $\Psi(s^*)=s^*$.

Step 1. $g(s)=\Lambda(1-s)[m+s\Delta/2]$ is a product of two factors non-negative on [0,1]: the first since s≤1; the second because it is linear and increasing in s, equal to m>0 at s=0 and to $m+\Delta/2=\gamma_{max}-1>0$ at s=1. So g(s)≥0 throughout, and likewise λ(s)=Λ(1−s)≥0 on [0,1]. Hence D(s)≥ρ>0 for every s∈[0,1], and v(s)=π/D(s) is continuous and strictly positive there.

Step 2. For any v>0, vΔ/(vΔ+ζ)∈(0,1) strictly, since ζ>0. Composed with v(s)>0 from Step 1, Ψ(s)∈(0, s^FB) for every s∈[0,1], and Ψ is continuous.

Step 3. Let Φ(s) ≡ Ψ(s)−s, continuous on [0,1]. By Step 2, Φ(0)=Ψ(0)>0, and Φ(1)=Ψ(1)−1<s^FB−1<0, using s^FB<1/2<1. By the intermediate value theorem, Φ(s*)=0 for some s*∈(0,1), and since Ψ(s*)∈(0, s^FB), so is s*.

A.3. A Sufficient Condition for Uniqueness

Lemma A.2. If

$$\frac{s^{FB}\Delta \cdot \pi[\theta\Lambda \cdot max(\Delta s^{FB}, \gamma_{max} - 1) + \Lambda]}{\zeta\rho^2} < 1 \quad \text{(A.1)}$$

the fixed point of (24) is unique.

Proof. g′(s)=Λ[(Δ/2−m)−sΔ] is linear and decreasing, so its extreme values on [0,1] occur at the endpoints: g′(0)=ΛΔs^FB and g′(1)=−Λ(γ$_{max}$−1), giving |g′(s)|≤Λ·max(Δs^FB, γ$_{max}$−1)≡$\bar{g}'$ throughout. Then |D′(s)|=|θg′(s)−Λ|≤θ$\bar{g}'$+Λ, and since v(s)=π/D(s) with D(s)≥ρ,

$$| v'(s) | = \frac{\pi | D'(s) |}{D^2} \leq \frac{\pi(\theta\bar{g}' + \Lambda)}{\rho^2}$$

Differentiating Ψ,

$$\Psi'(s) = \frac{s^{FB}\Delta\zeta \cdot v'(s)}{[v(s)\Delta + \zeta]^2},$$

$$| \Psi'(s) | \leq \frac{s^{FB}\Delta}{\zeta} | v'(s) | \leq \frac{s^{FB}\Delta\pi(\theta\bar{g}' + \Lambda)}{\zeta\rho^2}$$

$$| \Psi(s;\zeta) - s^{FB} | = s^{FB}\left(1 - \frac{v(s)\Delta}{v(s)\Delta + \zeta}\right) = \frac{s^{FB}\zeta}{v(s)\Delta + \zeta} \leq \frac{s^{FB}\zeta}{v_{\min}\Delta} \to 0 \; as \; \zeta \to 0$$

uniformly in s. Since s*(ζ)=Ψ(s*(ζ);ζ) for every ζ, |s*(ζ)−s^FB| ≤ sup$_s$|Ψ(s;ζ)−s^FB| → 0. Continuity of g then gives g(s*(ζ))→g(s^FB).

using [v(s)Δ+ζ]²≥ζ². If this bound is below one — condition (A.1) — Ψ is a contraction on [0,1], and the Banach fixed-point theorem gives a unique fixed point.

Condition (A.1) is sufficient but deliberately conservative, and it is not satisfied at the calibration in Section 5.6, where the left-hand side evaluates to approximately 12: screening cost would need to be roughly an order of magnitude higher, other parameters held fixed, before this bound alone could certify uniqueness. This is a property of the bound, not evidence against uniqueness. A direct numerical search over (0, s^FB) at the calibrated parameters finds exactly one root of (24), with Ψ′(s*)≈0.04 — far inside the locally stable region — and the same holds throughout the comparative-statics grids underlying Tables 3-5. A tighter, parameter-region-specific uniqueness argument is left for future work; it is not needed for any result in the main text, all of which concern properties holding at any fixed point (Propositions 1-2) or are verified directly by numerical search (Section 5.7).

A.4. The ζ→0 Limit

Lemma A.3. As ζ→0$^+$, every fixed point s*(ζ) of (24) converges to s^FB.

Proof. By Step 1 of A.2, g(s)≤g(s^FB) and λ(s)≤Λ for every s∈[0,1], so D(s)≤ρ+θg(s^FB)+Λ and hence v(s)≥v$_{min}$ ≡ π/[ρ+θg(s^FB)+Λ]>0, uniformly in s. Then

References

1. King R.G., Levine R. Finance and Growth: Schumpeter Might Be Right // Quarterly Journal of Economics. — 1993. — Vol. 108, No. 3. — P. 717-737.

2. Kortum S., Lerner J. Assessing the Contribution of Venture Capital to Innovation // RAND Journal of Economics. — 2000. — Vol. 31, No. 4. — P. 674-692.

3. Aghion P., Howitt P. A Model of Growth Through Creative Destruction // Econometrica. — 1992. — Vol. 60, No. 2. — P. 323-351.

4. Klette T.J., Kortum S. Innovating Firms and Aggregate Innovation // Journal of Political Economy. — 2004. — Vol. 112, No. 5. — P. 986-1018.

5. Aghion P., Akcigit U., Howitt P. What Do We Learn From Schumpeterian Growth Theory? // Handbook of Economic Growth. — Elsevier, 2014. — Vol. 2. — P. 515-563.

6. Acemoglu D., Akcigit U., Alp H., Bloom N., Kerr W.R. Innovation, Reallocation, and Growth // American Economic Review. — 2018. — Vol. 108, No. 11. — P. 3450-3491.

7. Aghion P., Howitt P., Mayer-Foulkes D. The Effect of Financial Development on Convergence: Theory and Evidence // Quarterly Journal of Economics. — 2005. — Vol. 120, No. 1. — P. 173-222.

8. Levine R. Finance and Growth: Theory and Evidence // Handbook of Economic Growth. — Elsevier, 2005. — Vol. 1. — P. 865-934.

9. Rajan R.G., Zingales L. Financial Dependence and Growth // American Economic Review. — 1998. — Vol. 88, No. 3. — P. 559-586.

10. Cecchetti S.G., Kharroubi E. Why Does Financial Sector Growth Crowd Out Real Economic Growth? // BIS Working Papers. — 2015. — No. 490.

11. Arcand J.-L., Berkes E., Panizza U. Too Much Finance? // Journal of Economic Growth. — 2015. — Vol. 20, No. 2. — P. 105-148.

12. Townsend R.M. Optimal Contracts and Competitive Markets with Costly State Verification // Journal of Economic Theory. — 1979. — Vol. 21, No. 2. — P. 265-293.

13. Diamond D.W. Financial Intermediation and Delegated Monitoring // Review of Economic Studies. — 1984. — Vol. 51, No. 3. — P. 393-414.

14. Holmström B., Tirole J. Financial Intermediation, Loanable Funds, and the Real Sector // Quarterly Journal of Economics. — 1997. — Vol. 112, No. 3. — P. 663-691.

15. Greenwood J., Sanchez J.M., Wang C. Financing Development: The Role of Information Costs // American Economic Review. — 2010. — Vol. 100, No. 4. — P. 1875-1891.

16. Kaplan S.N., Strömberg P. Venture Capitalists as Principals: Contracting, Screening, and Monitoring // American Economic Review. — 2001. — Vol. 91, No. 2. — P. 426-430.

17. Nanda R., Rhodes-Kropf M. Investment Cycles and Startup Innovation // Journal of Financial Economics. — 2013. — Vol. 110, No. 2. — P. 403-418.

18. Hsieh C.-T., Klenow P.J. Misallocation and Manufacturing TFP in China and India // Quarterly

Journal of Economics. — 2009. — Vol. 124, No. 4. — P. 1403-1448.

19. Restuccia D., Rogerson R. Policy Distortions and Aggregate Productivity with Heterogeneous Establishments // Review of Economic Dynamics. — 2008. — Vol. 11, No. 4. — P. 707-720.

20. Midrigan V., Xu D.Y. Finance and Misallocation: Evidence from Plant-Level Data // American Economic Review. — 2014. — Vol. 104, No. 2. — P. 422-458.

21. David J.M., Hopenhayn H.A., Venkateswaran V. Information, Misallocation, and Aggregate Productivity // Quarterly Journal of Economics. — 2016. — Vol. 131, No. 2. — P. 943-1005.

22. Peters M. Heterogeneous Markups, Growth, and Endogenous Misallocation // Econometrica. — 2020. — Vol. 88, No. 5. — P. 2037-2073.

23. Amirov A. Innovation Credit Expansion, Resource Misallocation, and the Quality of Economic Growth // Journal of Economic Growth and Social Welfare. — 2026. — No. 1. — P. 179–183.

24. De Loecker J., Eeckhout J., Unger G. The Rise of Market Power and the Macroeconomic Implications // Quarterly Journal of Economics. — 2020. — Vol. 135, No. 2. — P. 561-644.

FROM NUCLEAR SPECTRA TO PORTFOLIO RISK: RANDOM MATRIX THEORY AS A UNIFYING LANGUAGE FOR HIGH-DIMENSIONAL RISK ESTIMATION

Asgarov A.
ORCID: https://orcid.org/0009-0004-9672-0367
Department of Finance, Azerbaijan State University of Economics (UNEC), Baku, Azerbaijan
Cahangirzada R.
ORCID: https://orcid.org/0009-0008-2043-4179
DOI: 10.5281/zenodo.22107058

ABSTRACT

Random matrix theory (RMT) was introduced into physics as a means of making reliable, system-level statements about systems too complicated to solve in microscopic detail. Its subsequent migration into quantitative finance followed an almost identical epistemic logic. This review traces the conceptual arc that connects Wigner's statistical treatment of nuclear resonance spectra to the modern practice of covariance cleaning, portfolio optimisation, and high-dimensional risk estimation. We begin with the founding intuition of nuclear RMT—replacing an intractable deterministic Hamiltonian by an ensemble of matrices constrained only by symmetry—and develop the classical Gaussian ensembles, the semicircle law, level repulsion, Dyson's Coulomb-gas picture, and the principle of spectral universality. We then show how universality transferred the method, rather than the physics, into finance: the sample covariance matrix of asset returns plays the role of a noisy Hamiltonian, and the Marchenko–Pastur law supplies a null model for what pure estimation noise should look like in high dimensions. Building on this null model, we review the landmark econophysics results of the late 1990s, the market–sector–noise decomposition of correlation spectra, and the practical consequences for mean–variance optimisation, eigenvalue cleaning, linear and nonlinear shrinkage, and rotationally invariant estimators. We close with the principal limitations of the classical Marchenko–Pastur null—heavy tails, non-stationarity, and temporal dependence—and the research lines that address them: robust scatter estimation, heavy-tail-aware cleaning, free probability, dynamic spectral monitoring, and the broader convergence of finance-RMT with high-dimensional statistics and machine learning. The recurring lesson is methodological: RMT does not begin as a theory of markets but as a theory of what noise alone would produce in a large estimated covariance matrix, and that stance is what makes it durable.

Keywords: random matrix theory; Marchenko–Pastur law; covariance estimation; portfolio optimisation; eigenvalue cleaning; shrinkage estimators; econophysics; high-dimensional statistics.

1. Introduction

Random matrix theory entered physics as a way of saying something dependable about systems that were too complicated to solve exactly. Its migration into finance half a century later obeyed the same logic. In nuclear physics the object of interest was the many-body Hamiltonian of a heavy compound nucleus; in finance it became the empirical covariance or correlation matrix of many asset returns, estimated from limited and noisy data. In both settings the central move is identical. One stops trying to recover every microscopic interaction and instead asks whether the spectral statistics of an appropriately chosen random-matrix ensemble capture the robust, system-level structure that survives the complexity. That single conceptual continuity links Wigner's nuclear programme of the early 1950s to contemporary covariance cleaning, portfolio construction, and high-dimensional risk estimation in quantitative finance (Bouchaud & Potters, 2011; Bun et al., 2017).

This review is organised around that continuity rather than around a chronological catalogue of results. Section 2 reconstructs the origins of RMT in nuclear physics—the statistical hypothesis, the Gaussian ensembles, the semicircle law, level repulsion, and Dyson's Coulomb-gas picture. Section 3 isolates the idea of universality, which is the property that actually travels across disciplines. Section 4 builds the bridge to finance through the sample covariance matrix and the Marchenko–Pastur law. Section 5 reviews the landmark financial applications of the late 1990s and early 2000s. Section 6 turns to practice: why naive mean–variance optimisation is fragile in high dimensions, and how eigenvalue cleaning, shrinkage, and rotationally invariant estimators respond. Section 7 surveys the limitations of the classical null model and the research that now extends it. Throughout, the emphasis is on the methodological stance that finance imported, because that stance, more than any single formula, is what has proved durable.

2. Origins in Nuclear Physics

2.1 Wigner’s statistical hypothesis

Wigner's starting point was the statistical behaviour of resonance levels in compound nuclei. For heavy nuclei the underlying many-body interactions were far too complex to model in detail, yet the observed resonance spacings were irregular in a way that nonetheless suggested lawful statistics. In his 1951 work on the widths and spacings of nuclear resonance levels, Wigner argued that one could seek a statistical law of level spacings even when the exact Hamiltonian was out of reach (Wigner, 1951). This was the founding intuition of RMT: replace an intractable deterministic Hamiltonian by an ensemble of matrices carrying only the correct symmetry constraints, and study ensemble-robust spectral features instead of individual levels. Later reviews of nuclear RMT are explicit that this was not a shortcut for solving a particular nucleus but a genuinely new statistical description of spectral fluctuations in chaotic many-body systems (Guhr et al., 1998; Papenbrock & Weidenmüller, 2007; Weidenmüller & Mitchell, 2009).

2.2 The Gaussian ensembles and Dyson’s threefold way

The classical Gaussian ensembles formalised this idea. One writes a probability density on the space of matrices of the form

$$P(H) \propto exp[-a\beta \cdot Tr(H^2)],$$

with the matrix class fixed by symmetry: the Gaussian Orthogonal Ensemble (GOE) for real symmetric matrices, the Gaussian Unitary Ensemble (GUE) for complex Hermitian matrices, and the Gaussian Symplectic Ensemble (GSE) for quaternion self-dual matrices. Dyson's “threefold way” showed that these three ensembles correspond to the three fundamental symmetry classes governing time-reversal and spin, and Mehta's monograph turned the subject into a coherent theory with standard notation, methods, and results (Dyson, 1962d; Mehta, 2004). In the nuclear setting the symmetry class is not a technicality: it fixes the correct short-range level repulsion and the local spectral statistics.

2.3 The semicircle law and level repulsion

One of Wigner's earliest asymptotic results is the semicircle law. For a suitably normalised large Wigner matrix, the empirical eigenvalue density converges to

$$\rho\_sc(\lambda) = (1 / 2\pi\sigma^2) \cdot \sqrt{(4\sigma^2 - \lambda^2)}, for\ |\lambda| \leq 2\sigma,$$

and vanishes outside this support (Wigner, 1955, 1958). The semicircle describes the global density of states. A different but equally important observable is the local spacing between adjacent unfolded eigenvalues. Wigner's surmise for the GOE is

$$P\_GOE(s) \approx (\pi/2)\, s \cdot exp(-\pi\, s^2 / 4),$$

which already exhibits the crucial phenomenon of level repulsion, since P(s) → 0 as s → 0. More generally, for the three classical symmetry classes one has P_β(s) ~ s^β near the origin with β = 1, 2, 4, so that exact degeneracies are statistically suppressed unless a symmetry forces them. Level repulsion is the qualitative fingerprint that distinguishes a complex, “chaotic” spectrum from a sequence of independent levels.

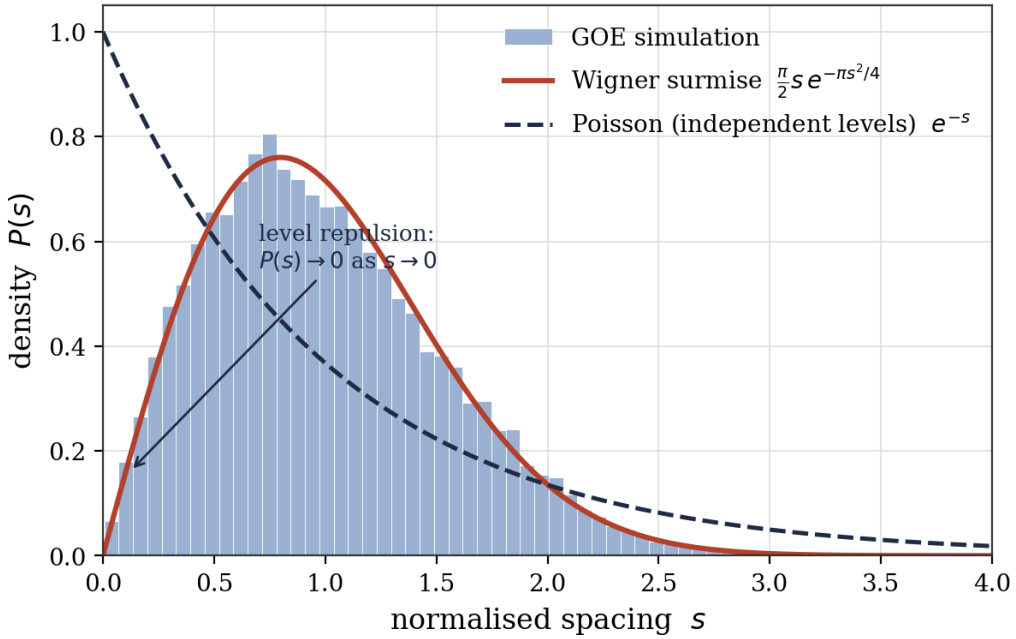


Figure 1. Nearest-neighbour level spacings from a simulated Gaussian Orthogonal Ensemble (eight matrices, N = 4000), unfolded to unit mean spacing. The histogram follows the Wigner surmise rather than the Poisson law expected for independent levels; the vanishing of P(s) as s → 0 is the level repulsion characteristic of complex spectra. Figure generated by the author.

2.4 Dyson’s Coulomb-gas picture

Dyson's deeper contribution was to recast the eigenvalues as an interacting gas. The joint eigenvalue density for the Gaussian ensembles can be written as

$$P(\lambda_1, \ldots, \lambda_N) \propto \prod\_\{i<j\} |\lambda\_i - \lambda\_j|^\beta \cdot exp(-c \cdot \Sigma\_i\, \lambda\_i^2),$$

which is equivalent to a one-dimensional Coulomb gas with logarithmic repulsion and quadratic confinement (Dyson, 1962a, 1962b, 1962c). This representation makes level repulsion almost visually obvious: the Vandermonde factor $\prod\_\{i<j\} |\lambda\_i - \lambda\_j|^\beta$ penalises configurations in which two eigenvalues approach one another. Dyson's 1962 sequence of papers and Mehta's later synthesis established the conceptual core that finance would eventually import—namely that the information of interest lives in the geometry of eigenvalues and eigenvectors, not in any literal physical interpretation of the matrix entries (Mehta, 2004).

3. Universality and the Statistical Philosophy of RMT

The single most transferable idea in RMT is universality. In modern language, many local spectral statistics are largely insensitive to the fine details of the matrix-element distribution once the symmetry class and the large-size scaling are fixed. In the original nuclear application this meant that the exact microscopic interaction inside a complicated nucleus could vary considerably, yet the spacing statistics still followed the same ensemble law. Reviews in quantum chaos and nuclear physics stress exactly this point: the theory does not predict the detailed spectrum of one nucleus, but rather the fluctuation pattern expected generically of any complex system sharing the same symmetries (Guhr et al., 1998; Weidenmüller & Mitchell, 2009).

This philosophical shift is what made the move into finance possible at all. Financial markets are not quantum systems, and asset-return matrices are not nuclear Hamiltonians. But once one accepts that the object of interest is not a literal physical operator but rather a large matrix whose collective spectral statistics arise from complexity, weak information, and finite-sample noise, the same RMT logic becomes available. Finance imported the methodological stance rather than the nuclear interpretation: use a random-matrix null model to separate statistically universal noise from structurally meaningful deviations (Bun et al., 2017). Universality is what licenses this transfer, because it guarantees that the null model's predictions do not depend on the unrealistic assumption that returns are exactly Gaussian.

4. The Bridge to Finance: Sample Covariance and the Marchenko–Pastur Law

Let r_{it} denote the return of asset i = 1, …, N observed over t = 1, …, T. Arranging the returns in an N × T matrix R, the sample covariance matrix is

$$S = (1/T) \cdot R\,R^{T},$$

and the sample correlation matrix follows by standardising the variances,

$$C = D^{-\frac{1}{2}} \cdot S \cdot D^{-\frac{1}{2}},$$
$$with\ D = diag(S_{11}, \ldots, S\_NN).$$

In practice both N and T are large and of comparable order, so the empirical spectrum of C is strongly distorted by estimation noise even when the true correlations are weak or absent. This is the financial analogue of Wigner's dilemma: the object one actually wants—the "true" covariance or correlation structure—is obscured by a large random component arising purely from finite sampling.

The appropriate null model for the eigenvalue density of a large sample covariance matrix is the Marchenko–Pastur (MP) law (Marchenko & Pastur, 1967). If the true covariance is σ²I, the entries are independent with finite variance, and N, T → ∞ with fixed aspect ratio

$$q = N / T,$$

then the limiting eigenvalue density is

$$\rho\_MP(\lambda) = (1 / 2\pi\, q\, \sigma^2\, \lambda) \cdot \sqrt{[(\lambda_+ - \lambda)(\lambda - \lambda_-)]},$$
$$for\ \lambda \in [\lambda_-, \lambda_+],$$

with spectral edges

$$\lambda\pm = \sigma^2\,(1 \pm \sqrt{q})^2.$$

For a correlation matrix one typically sets σ² = 1. When q > 1 there is an additional point mass at zero of weight 1 − 1/q. Much of the finance literature uses the inverse convention Q = T/N = 1/q, in which the edges become λ± = σ² (1 ± Q^(−1/2))². The larger q is, the wider the purely noisy band becomes; as q → 0 the band collapses and estimation noise becomes spectrally negligible.

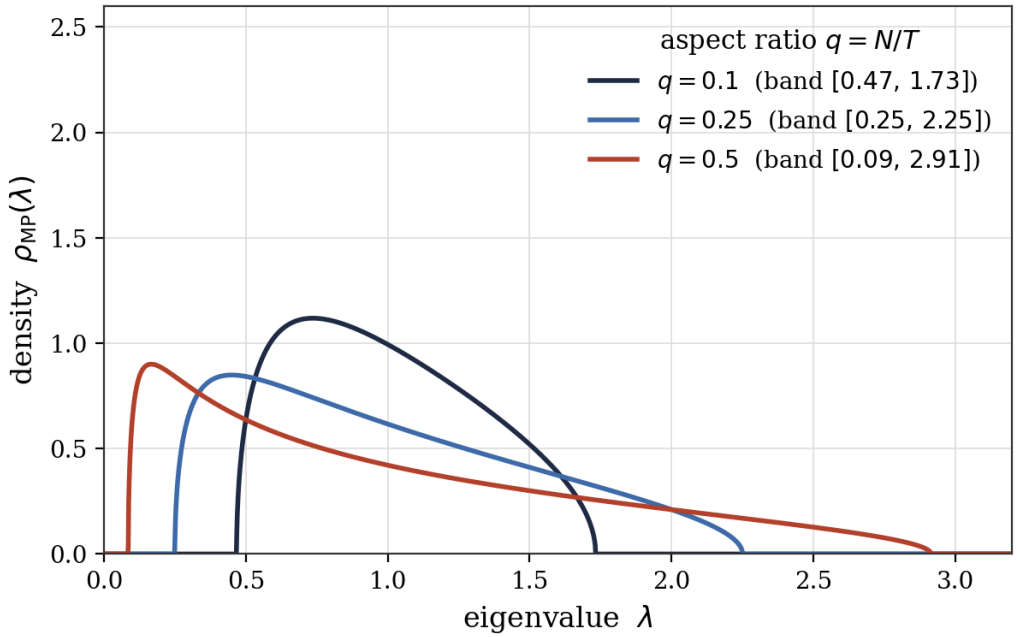


Figure 2. The Marchenko–Pastur density for three aspect ratios q = N/T. As q grows, the noise band [λ−, λ+] widens and a larger share of the spectrum is consistent with pure estimation noise; as q → 0 the band collapses. Figure generated by the author.

This is precisely the point at which nuclear-physics RMT became financially useful. The MP law supplies a null spectrum: if the empirical eigenvalue density of a correlation matrix lies inside the MP band, one should first suspect sampling noise rather than genuine economic structure. Eigenvalues that systematically sit outside the band are candidates for real factors. In finance, therefore, RMT does not begin as a theory of markets at all; it begins as a theory of what noise alone would look like in a large estimated covariance matrix, and only then does the residual—the part the null cannot explain—become the object of economic interpretation.

5. Landmark Financial Applications

5.1 Noise dressing of financial correlation matrices

The decisive breakthrough came in the late 1990s. Laloux, Cizeau, Bouchaud, and Potters (1999) compared empirical stock-return correlation spectra against the random-matrix null and found that, for the S&P 500

data they studied, roughly 94% of the eigenvalues fell within the band predicted by the random-matrix benchmark. Their interpretation was stark: most of the empirical spectrum was dominated by noise, and only a small tail of outlying eigenvalues carried economically meaningful information. In a closely related paper they framed RMT explicitly as a method for extracting the relevant correlations needed for risk management and portfolio construction (Laloux et al., 2000). The phrase "noise dressing" captured the central message: the empirical correlation matrix is a noisy dressing of a much lower-dimensional signal.

5.2 The market–sector–noise decomposition

Almost simultaneously, Plerou, Gopikrishnan, Rosenow, Amaral, and Stanley (1999) published their analysis of universal and non-universal properties of cross-correlations in financial time series, followed by a more detailed treatment that became a standard reference (Plerou et al., 2002; see also Plerou et al., 2001). Their central result was the now-canonical spectral decomposition of financial correlation matrices: a broad MP-like bulk interpreted as noise, a single largest eigenvalue reflecting a collective market-wide mode, and a handful of further outliers whose eigenvectors localise on sectoral or industry groups. RMT thereby converted the correlation matrix from an undifferentiated object into a hierarchy of modes—market, sector, and noise.

The "market mode" interpretation proved especially influential. The leading eigenvector is typically close to a same-sign vector across all assets, so it loads on the common co-movement of the whole market. The next informative eigenvectors are more localised and are often interpretable as industry or sector structure, although the precise mapping depends on the market, the sampling window, and the level of aggregation. That interpretation was robust enough to become part of the standard vocabulary of quantitative finance.

Figure 3 illustrates this market–sector–noise structure on a synthetic dataset built from a single dominant market factor, four weaker sector factors, and idiosyncratic noise. The dimensions (N = 120 assets, T = 480 days, giving an aspect ratio q = N/T = 0.25) are chosen to reflect a realistic estimation regime in which the sample length is comparable to the number of assets—large enough for the Marchenko–Pastur band to be clearly resolved, yet small enough for finite-sample noise to dominate the bulk of the spectrum.

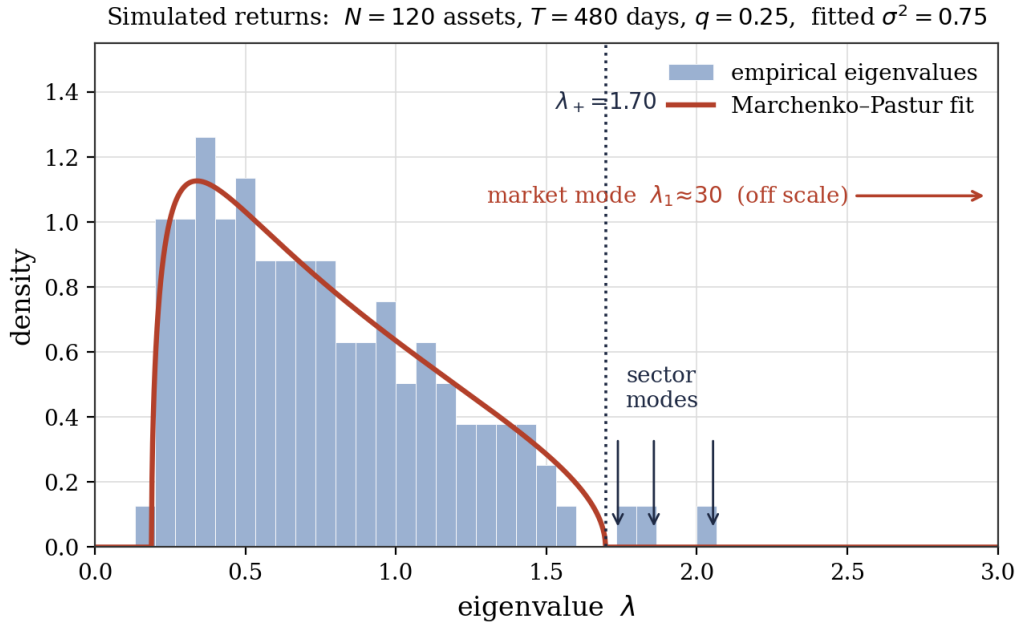


Figure 3. Eigenvalue spectrum of a correlation matrix estimated from simulated returns (N = 120 assets, T = 480 days, q = 0.25) built from one strong market factor, four weaker sector factors, and idiosyncratic noise. About 97% of the eigenvalues fall inside the fitted Marchenko–Pastur band; a single large eigenvalue (the market mode, $\lambda_1 \approx 30$, off scale) and a few sector modes lie just beyond λ_+. The variance carried by the deviating modes is removed before fitting ($\sigma^2 \approx 0.75$), following Laloux et al. (2000). Figure generated by the author.

5.3 Cross-market evidence

Subsequent studies extended the same spectral logic well beyond US equities. Utsugi, Ino, and Oshikawa (2004) applied RMT to Japanese equities, while work on emerging markets repeatedly found the same broad separation between a dominant market factor and a much noisier residual bulk, though often with weaker or less sharply resolved sector modes than in large developed markets. This comparative dimension mattered conceptually: RMT did not merely filter noise, it furnished a common language for describing how market structure itself varies across countries and stages of financial development.

5.4 The econophysics synthesis

Bouchaud and Potters' broader programme then gave the field its durable synthesis. Their textbook integrated cross-correlation matrices, estimation noise, and portfolio implications into a unified econophysics framework (Bouchaud & Potters, 2003), and their later handbook chapter surveyed MP-based null models, eigenvalue cleaning, free-probability tools, and out-of-sample risk estimation (Bouchaud & Potters, 2011). By that stage RMT in finance had become less a novelty

than a standard component of the toolkit for high-dimensional covariance estimation, a status consolidated by the comprehensive review of Bun, Bouchaud, and Potters (2017) and the later pedagogical treatment of Potters and Bouchaud (2021).

6. Practical Impact on Portfolio Construction and Risk Management

6.1 The instability of mean–variance optimisation

The practical stakes are clearest in mean–variance optimisation (Markowitz, 1952). In the minimum-variance problem with only the budget constraint $1^{T}w = 1$, the optimal weights are

$$w\_min \;=\; (\Sigma^{-1}\,1) \,/\, (1^{T}\,\Sigma^{-1}\,1).$$

This formula is compact but notoriously unstable, because it requires inverting an estimated covariance matrix. Errors in the small empirical eigenvalues of Σ are magnified by Σ^{-1}, so the optimiser overreacts to noisy directions and produces extreme, poorly diversified weights. Michaud's (1989) critique of the "Markowitz optimisation enigma" predated the finance-RMT literature, but RMT later supplied a precise spectral diagnosis of why naive optimisation is so fragile in high dimensions: the directions the optimiser trusts most are exactly the ones the MP null identifies as least reliable.

6.2 Eigenvalue cleaning

The earliest RMT-based cleaning schemes were operationally simple. One diagonalises the empirical covariance or correlation matrix,

$$C \;=\; \Sigma\_\{i=1\}\^\{N\}\, \lambda\_i\, u\_i\, u\_i^{T},$$

identifies the set $\mathcal{N}$ of eigenvalues lying inside the MP noise band, and replaces those noisy eigenvalues by a cleaned value while preserving the eigenvectors,

$$C_{clean} \;=\; \Sigma_{\{i\notin\mathcal{N}\}}\lambda_i u_i u_{i^{T}} \;+$$
$$+\Sigma\_\{i\in\mathcal{N}\}\;\lambda\_noise\; u\_i\, u\_i^{T}.$$

The simplest implementations either clip the noisy eigenvalues to their common average or reconstruct only the few informative modes. Applied work argued that such cleaning improves conditioning and stabilises optimised portfolios (Laloux et al., 2000; Plerou et al., 2002). Pafka and Kondor (2002, 2003), however, already flagged an important nuance: the gains depend on the specific portfolio problem, the sample ratio T/N, and the presence of real-world constraints, so cleaning is not a guaranteed improvement in every regime.

6.3 Comparison with shrinkage estimators

This is where the comparison with shrinkage methods becomes essential. Linear shrinkage in the sense of Ledoit and Wolf (2004) replaces the sample covariance S by a convex combination,

$$\Sigma\_LW \;=\; \delta\, F \;+\; (1-\delta)\, S,$$

where F is a structured target—such as a multiple of the identity or a factor-based matrix—and $\delta \in [0, 1]$ is chosen optimally from the data. The objective resembles that of RMT cleaning, namely to reduce estimation error and improve conditioning, but the mechanism differs. Linear shrinkage pulls the whole matrix towards a target, whereas classical eigenvalue cleaning leaves the eigenvectors unchanged and regularises primarily through the spectrum. Both are responses to the same high-dimensional pathology, and conceptually they overlap: RMT cleaning can be read as a data-driven, spectrally structured form of shrinkage.

6.4 Nonlinear shrinkage and rotationally invariant estimators

Later work sharpened this comparison considerably. Ledoit and Péché (2011) analysed the asymptotic relationship between sample and population eigenvectors, which opened the way to nonlinear shrinkage: instead of shrinking every eigenvalue by the same factor, one shrinks each sample eigenvalue by an amount derived from its asymptotically optimal bias correction. Ledoit and Wolf (2012, 2020) turned this into implementable estimators with strong finite-sample performance. Relative to simple MP clipping, nonlinear shrinkage is less ad hoc and typically more accurate, precisely because it recognises that different parts of the spectrum should be regularised by different amounts.

The Bun–Bouchaud–Potters line of work pushed the same idea further with rotationally invariant estimators (RIEs). In the absence of trustworthy prior information about the true eigenvectors, a rotationally invariant estimator keeps the empirical eigenvectors u_i but replaces the empirical eigenvalues λ_i by optimally cleaned values ξ_i,

$$C\_RIE \;=\; \Sigma\_\{i=1\}\^\{N\}\, \xi\_i\, u\_i\, u\_i^{T}.$$

For multiplicative noise, the asymptotic oracle estimator can be written in the form

$$\xi(\lambda) \;=\; \lambda \,/\, |1 - q + q\,\lambda\cdot g\_E(\lambda - i0^{+})|^{2},$$

where g_E is the limiting Stieltjes transform of the empirical spectrum. The operational point matters more than the formula: an RIE produces a nonlinear, eigenvalue-by-eigenvalue correction derived from RMT itself rather than from a fixed shrinkage target. Bun et al. (2016, 2017) argued that such estimators perform especially well for the large, noisy correlation matrices that arise in finance.

6.5 Empirical caveats

Empirical comparative work reinforces the message but tempers it with caveats. Pantaleo, Tumminello, Lillo, and Mantegna (2011) compared nine covariance estimators and found that improved estimators—both RMT-based schemes and shrinkage procedures—can substantially reduce realised portfolio risk when T/N is close to one and short selling is allowed. When short selling is prohibited and T > N, however, the plain sample covariance is harder to beat on realised risk, even though improved estimators can still deliver more diversified weights. This finding is important for a balanced review because it forestalls overstatement: RMT cleaning is not a universal dominance result but a strong response to a well-defined high-dimensional regime.

6.6 Broader uses in risk management

Beyond portfolio weights, the same spectral logic informs risk management more broadly. Large-dimensional covariance matrices are central to stress testing, risk decomposition, risk parity, and factor attribution. Recent work on large dynamic covariance matrices shows how high-dimensional regularisation and shrinkage can be integrated with time-varying volatility models and intraday information (De Nard et al., 2022), while RMT-based approaches to credit risk replace a fixed correlation matrix by an ensemble of random correlations in order to quantify the limits of diversification under uncertain dependence (Münnix et al., 2014).

In market-state monitoring, researchers track the leading eigenvalues and related spectral observables as indicators of regime shifts or rising systemic fragility (Jurczyk et al., 2017).

7. Limitations and Recent Developments

7.1 Beyond the Marchenko–Pastur null

The pure MP null model is analytically elegant and often empirically useful, but it rests on assumptions that financial data routinely violate. Daily and intraday returns are heavy-tailed, heteroskedastic, and non-stationary; correlations evolve over time; and temporal dependence contaminates the independence assumption behind the classical Wishart setting. El Karoui (2009) showed that in high dimensions the spectral behaviour of large covariance and correlation matrices can change materially under elliptical and related distributions, so that MP-type equations must be generalised. The original finance use of the MP law is therefore best understood as a first null model, not a complete probabilistic description of market data.

7.2 Robust scatter estimation

This limitation motivated a robust-RMT literature. Couillet, Pascal, and Silverstein (2015) studied high-dimensional M-estimators, such as Maronna's robust scatter estimator, in the random-matrix regime with elliptically distributed samples, and Zhang, Cheng, and Singer (2016) established Marchenko–Pastur-type behaviour for suitably scaled Tyler-type estimators. The financial implication is direct: to preserve the spectral-denoising benefits of RMT while taking fat tails seriously, one often has to replace the sample covariance matrix itself by a robust scatter estimator before applying spectral cleaning.

7.3 Heavy-tail-aware cleaning

A parallel development moves from Gaussian-style cleaning to genuinely heavy-tail-aware cleaning. Bongiorno and Berritta (2023) develop optimal covariance cleaning for heavy-tailed distributions using information-theoretic arguments, pushing the Bun–Bouchaud–Potters programme into a regime much closer to real financial return data than the original Wishart benchmark. This line of work addresses one of the most frequent criticisms of early finance-RMT papers—that Gaussian assumptions are too restrictive for real markets—without abandoning the spectral-cleaning philosophy.

7.4 Free probability and spectral transforms

A second major line since roughly 2015 is the growing role of free probability and related spectral transforms. In finance, free-convolution methods allow one to deconvolve noise and signal spectra, to analyse additive and multiplicative perturbations, and to derive estimators without committing to a low-dimensional factor model. Bun et al. (2017) present free probability as one of the central analytic tools behind modern covariance cleaning, and Bouchaud and Potters (2011) place it alongside MP theory, random singular-value analysis, and eigenvector dynamics as part of the post-1999 maturation of the field.

7.5 From static to dynamic spectral analysis

A third development is the shift from static to dynamic spectral analysis. In real markets both eigenvalues and eigenvectors drift, with pronounced changes during crises. Large dynamic covariance models, rolling-window RMT diagnostics, and early-warning applications therefore attempt to detect when the leading eigenmodes strengthen, merge, or destabilise. The work of Jurczyk et al. (2017) on critical transitions belongs to this literature: it connects systemic-risk-style behaviour to changes in the empirical spectrum rather than to any single balance-sheet channel. This is close in spirit to the original RMT ethos—diagnose structure from collective spectral statistics rather than from a fully specified microscopic model.

7.6 New markets

Applications have also spread far beyond US equities, into Japanese, Chinese, Indian, and other equity markets, into credit-risk models where random correlations yield conservative loss estimates, and into newer markets such as cryptocurrencies. The crypto case is instructive because it tends to accentuate the limitations of the classical null. Studies find a strong market mode and clear collective structure, but also more pronounced deviations from standard equity-market regularities, consistent with stronger non-stationarity and a different microstructure (Stošić et al., 2018; Drożdż et al., 2025).

7.7 Convergence with machine learning and high-dimensional statistics

Finally, there is a methodological convergence with modern machine learning and high-dimensional statistics. Many of the problems RMT helps solve in finance—covariance regularisation, denoised principal component analysis, signal-versus-noise separation, and stability in large dimensions—are also central to machine learning. The recent RMT-for-machine-learning literature does not displace the finance work; rather it shows that finance was an early and important test bed for ideas that now sit inside a much broader theory of high-dimensional estimation (Couillet & Liao, 2022). In that sense, finance helped transform RMT from a specialised branch of mathematical physics into a general-purpose language for noisy spectral inference.

8. Conclusion

The path from nuclear spectra to portfolio risk is best read not as an analogy between two unrelated fields but as the reuse of a single methodological stance. Wigner's insight was that when a system is too complex to model microscopically, one can still characterise the robust statistics of its spectrum by replacing the true operator with a symmetry-constrained random ensemble. Universality guaranteed that those statistics did not depend on microscopic detail, and that guarantee is exactly what allowed the method to cross into finance, where the “operator” is a noisy sample covariance matrix rather than a Hamiltonian. The Marchenko–Pastur law gave finance its null spectrum; the late-1990s econophysics papers used that null to separate market and sector modes from noise; and the subsequent literature on cleaning, shrinkage, and rotationally invariant estimators turned the insight into deployable estimators for portfolio construction and risk management.

The honest summary is that finance-RMT is powerful but bounded. Its classical null rests on assumptions—Gaussianity, independence, stationarity—that

real markets violate, and the empirical evidence shows that cleaning helps decisively in some regimes and only marginally in others. The most productive recent work accepts this and extends the framework: robust scatter estimation for heavy tails, information-theoretic cleaning, free-probability transforms, and dynamic spectral monitoring. What unifies all of it is the founding ethos inherited from nuclear physics—diagnose system-level structure from collective spectral statistics rather than from a fully specified microscopic model. That ethos, rather than any single equation, is the enduring contribution of random matrix theory to quantitative finance.

Acknowledgements

The authors thank Anar Muradov, PhD, Senior Lecturer (ORCID: 0000-0002-5259-4424), for helpful discussions and for constructive comments on an earlier draft of this manuscript. Any remaining errors are the sole responsibility of the authors.

References

1. Bongiorno, C., & Berritta, M. (2023). Optimal covariance cleaning for heavy-tailed distributions: Insights from information theory. Physical Review E, 108(5), 054133. https://doi.org/10.1103/PhysRevE.108.054133

2. Bouchaud, J.-P., & Potters, M. (2003). Theory of financial risk and derivative pricing: From statistical physics to risk management (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511753893

3. Bouchaud, J.-P., & Potters, M. (2011). Financial applications of random matrix theory: A short review. In G. Akemann, J. Baik, & P. Di Francesco (Eds.), The Oxford handbook of random matrix theory (pp. 824–850). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198744191.013.40

4. Bun, J., Allez, R., Bouchaud, J.-P., & Potters, M. (2016). Rotational invariant estimator for general noisy matrices. IEEE Transactions on Information Theory, 62(12), 7475–7490. https://doi.org/10.1109/TIT.2016.2616132

5. Bun, J., Bouchaud, J.-P., & Potters, M. (2017). Cleaning large correlation matrices: Tools from random matrix theory. Physics Reports, 666, 1–109. https://doi.org/10.1016/j.physrep.2016.10.005

6. Couillet, R., & Liao, Z. (2022). Random matrix methods for machine learning. Cambridge University Press. https://doi.org/10.1017/9781009128490

7. Couillet, R., Pascal, F., & Silverstein, J. W. (2015). The random matrix regime of Maronna's M-estimator with elliptically distributed samples. Journal of Multivariate Analysis, 139, 56–78. https://doi.org/10.1016/j.jmva.2015.02.020

8. De Nard, G., Engle, R. F., Ledoit, O., & Wolf, M. (2022). Large dynamic covariance matrices: Enhancements based on intraday data. Journal of Banking & Finance, 138, 106426. https://doi.org/10.1016/j.jbankfin.2022.106426

9. Drożdż, S., Jarosz, P., Kwapień, J., Skupień, M., & Wątorek, M. (2025). Detrended cross-correlations and their random matrix limit: An example from the cryptocurrency market. Entropy, 27(12), 1236. https://doi.org/10.3390/e27121236

10. Dyson, F. J. (1962a). Statistical theory of the energy levels of complex systems. I. Journal of Mathematical Physics, 3(1), 140–156. https://doi.org/10.1063/1.1703773

11. Dyson, F. J. (1962b). Statistical theory of the energy levels of complex systems. II. Journal of Mathematical Physics, 3(1), 157–165. https://doi.org/10.1063/1.1703774

12. Dyson, F. J. (1962c). Statistical theory of the energy levels of complex systems. III. Journal of Mathematical Physics, 3(1), 166–175. https://doi.org/10.1063/1.1703775

13. Dyson, F. J. (1962d). The threefold way: Algebraic structure of symmetry groups and ensembles in quantum mechanics. Journal of Mathematical Physics, 3(6), 1199–1215. https://doi.org/10.1063/1.1703863

14. El Karoui, N. (2009). Concentration of measure and spectra of random matrices: Applications to correlation matrices, elliptical distributions and beyond. The Annals of Applied Probability, 19(6), 2362–2405. https://doi.org/10.1214/08-AAP548

15. Guhr, T., Müller-Groeling, A., & Weidenmüller, H. A. (1998). Random-matrix theories in quantum physics: Common concepts. Physics Reports, 299(4–6), 189–425. https://doi.org/10.1016/S0370-1573(97)00088-4

16. Jurczyk, J., Rehberg, T., Eckrot, A., & Morgenstern, I. (2017). Measuring critical transitions in financial markets. Scientific Reports, 7, 11564. https://doi.org/10.1038/s41598-017-11854-1

17. Laloux, L., Cizeau, P., Bouchaud, J.-P., & Potters, M. (1999). Noise dressing of financial correlation matrices. Physical Review Letters, 83(7), 1467–1470. https://doi.org/10.1103/PhysRevLett.83.1467

18. Laloux, L., Cizeau, P., Bouchaud, J.-P., & Potters, M. (2000). Random matrix theory and financial correlations. International Journal of Theoretical and Applied Finance, 3(3), 391–397. https://doi.org/10.1142/S0219024900000255

19. Ledoit, O., & Péché, S. (2011). Eigenvectors of some large sample covariance matrix ensembles. Probability Theory and Related Fields, 151(1–2), 233–264. https://doi.org/10.1007/s00440-010-0298-3

20. Ledoit, O., & Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. Journal of Multivariate Analysis, 88(2), 365–411. https://doi.org/10.1016/S0047-259X(03)00096-4

21. Ledoit, O., & Wolf, M. (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. The Annals of Statistics, 40(2), 1024–1060. https://doi.org/10.1214/12-AOS989

22. Ledoit, O., & Wolf, M. (2020). Analytical nonlinear shrinkage of large-dimensional covariance matrices. The Annals of Statistics, 48(5), 3043–3065. https://doi.org/10.1214/19-AOS1921

23. Marchenko, V. A., & Pastur, L. A. (1967). Distribution of eigenvalues for some sets of random matrices. Mathematics of the USSR-Sbornik, 1(4), 457–483. https://doi.org/10.1070/SM1967v001n04ABEH001994

24. Markowitz, H. (1952). Portfolio selection. The Journal of Finance, 7(1), 77–91. https://doi.org/10.1111/j.1540-6261.1952.tb01525.x

25. Mehta, M. L. (2004). Random matrices (3rd ed.). Pure and Applied Mathematics (Vol. 142). Elsevier.

26. Michaud, R. O. (1989). The Markowitz optimization enigma: Is "optimized" optimal? Financial Analysts Journal, 45(1), 31–42. https://doi.org/10.2469/faj.v45.n1.31

27. Münnix, M. C., Schäfer, R., & Guhr, T. (2014). A random matrix approach to credit risk. PLOS ONE, 9(5), e98030. https://doi.org/10.1371/journal.pone.0098030

28. Pafka, S., & Kondor, I. (2002). Noisy covariance matrices and portfolio optimization. The European Physical Journal B, 27(2), 277–280. https://doi.org/10.1140/epjb/e20020153

29. Pafka, S., & Kondor, I. (2003). Noisy covariance matrices and portfolio optimization II. Physica A: Statistical Mechanics and Its Applications, 319, 487–494. https://doi.org/10.1016/S0378-4371(02)01499-1

30. Pantaleo, E., Tumminello, M., Lillo, F., & Mantegna, R. N. (2011). When do improved covariance matrix estimators enhance portfolio optimization? An empirical comparative study of nine estimators. Quantitative Finance, 11(7), 1067–1080. https://doi.org/10.1080/14697688.2010.534813

31. Papenbrock, T., & Weidenmüller, H. A. (2007). Colloquium: Random matrices and chaos in nuclear spectra. Reviews of Modern Physics, 79(3), 997–1013. https://doi.org/10.1103/RevModPhys.79.997

32. Plerou, V., Gopikrishnan, P., Rosenow, B., Amaral, L. A. N., & Stanley, H. E. (1999). Universal and nonuniversal properties of cross correlations in financial time series. Physical Review Letters, 83(7), 1471–1474.
https://doi.org/10.1103/PhysRevLett.83.1471

33. Plerou, V., Gopikrishnan, P., Rosenow, B., Amaral, L. A. N., & Stanley, H. E. (2001). Collective behavior of stock price movements—A random matrix theory approach. Physica A: Statistical Mechanics and Its Applications, 299(1–2), 175–180. https://doi.org/10.1016/S0378-4371(01)00293-X

34. Plerou, V., Gopikrishnan, P., Rosenow, B., Amaral, L. A. N., Guhr, T., & Stanley, H. E. (2002). Random matrix approach to cross correlations in financial data. Physical Review E, 65(6), 066126. https://doi.org/10.1103/PhysRevE.65.066126

35. Potters, M., & Bouchaud, J.-P. (2021). A first course in random matrix theory: For physicists, engineers and data scientists. Cambridge University Press. https://doi.org/10.1017/9781108768900

36. Stošić, D., Stošić, D., Ludermir, T. B., & Stošić, T. (2018). Collective behavior of cryptocurrency price changes. Physica A: Statistical Mechanics and Its Applications, 507, 499–509. https://doi.org/10.1016/j.physa.2018.05.050

37. Utsugi, A., Ino, K., & Oshikawa, M. (2004). Random matrix theory analysis of cross correlations in financial markets. Physical Review E, 70(2), 026110. https://doi.org/10.1103/PhysRevE.70.026110

38. Weidenmüller, H. A., & Mitchell, G. E. (2009). Random matrices and chaos in nuclear physics: Nuclear structure. Reviews of Modern Physics, 81(2), 539–589. https://doi.org/10.1103/RevModPhys.81.539

39. Wigner, E. P. (1951). On the statistical distribution of the widths and spacings of nuclear resonance levels. Mathematical Proceedings of the Cambridge Philosophical Society, 47(4), 790–798. https://doi.org/10.1017/S0305004100027237

40. Wigner, E. P. (1955). Characteristic vectors of bordered matrices with infinite dimensions. Annals of Mathematics, 62(3), 548–564. https://doi.org/10.2307/1970079

41. Wigner, E. P. (1958). On the distribution of the roots of certain symmetric matrices. Annals of Mathematics, 67(2), 325–327. https://doi.org/10.2307/1970008

42. Zhang, T., Cheng, X., & Singer, A. (2016). Marčenko–Pastur law for Tyler's M-estimator. Journal of Multivariate Analysis, 149, 114–123. https://doi.org/10.1016/j.jmva.2016.03.010

СОВРЕМЕННОЕ СОСТОЯНИЕ И ПЕРСПЕКТИВЫ РАЗВИТИЯ ЦИФРОВОЙ ЭКОНОМИКИ КАЗАХСТАНА

Ибраева Э.А.
доктор DBA, профессор, «Казахстанский университет инновационных и телекоммуникационных систем», г. Уральск, Казахстан
Ибраев А.Т.
доктор физико-математических наук, профессор, «Казахстанская Академия информации и бизнеса», г. Алматы, Казахстан
Робинсон А.Д.
бакалавр, младший научный сотрудник, «Казахстанская Академия информации и бизнеса», г. Алматы, Казахстан

CURRENT STATE AND DEVELOPMENT PROSPECTS OF KAZAKHSTAN'S DIGITAL ECONOMY

Ibrayeva E.
DBA, Professor, Kazakhstan University of Innovation and Telecommunication Systems, Uralsk, Kazakhstan
Ibrayev A.
Doctor of Physical and Mathematical Sciences, Professor, Kazakhstan Academy of Information and Business, Almaty, Kazakhstan
Robinson A.
Bachelor of Science, Junior Researcher, Kazakhstan Academy of Information and Business, Almaty, Kazakhstan
DOI: 10.5281/zenodo.22107088

АННОТАЦИЯ

В современном мире цифровая экономика является важнейшим фактором устойчивого развития, повышения конкурентоспособности и успешной интеграции стран в глобальное экономическое пространство. Активное внедрение новых цифровых технологий и искусственного интеллекта с использованием Больших данных (Big Data) и облачных вычислений трансформирует и совершенствует традиционные отрасли экономики, ускоряет процесс формирования новых возможностей для развития региональных и глобальных рынков. Казахстан в настоящее время является страной с развивающейся экономикой, которая активно реализует стратегию цифровой трансформации, стремится занять лидирующие позиции в Центральной Азии и укрепить свою роль в глобальной цифровой системе.

В данной работе проводится сравнительный анализ уровня цифровизации экономики Казахстана с государствами Центральной Азии, Россией, странами Европейского союза (ЕС) и странами, которые являются мировыми лидерами в области цифровой экономики. Кроме того, в статье проводится анализ основных организационных и теоретических проблем цифровизации экономики. На основе изучения методологии цифровизации и сравнения уровня цифровой экономики Казахстана с достижениями других стран мировой экономики в статье предлагается перечень основных необходимых мероприятий для ускоренного выхода Казахстана на самые передовые уровни развития инновационной экономики.

ABSTRACT

In today's world, the digital economy is a critical factor in sustainable development, increased competitiveness, and the successful integration of countries into the global economic space. The active implementation of new digital technologies and artificial intelligence using Big Data and cloud computing is transforming and improving traditional economic sectors, accelerating the emergence of new opportunities for the development of regional and global markets. Kazakhstan is currently a developing economy actively implementing a digital transformation strategy, striving to become a leader in Central Asia, and strengthen its role in the global digital system.

This paper provides a comparative analysis of the level of digitalization of Kazakhstan's economy with that of Central Asian countries, Russia, the European Union (EU), and countries that are global leaders in the digital economy. Furthermore, the article analyzes the main organizational and theoretical challenges of economic digitalization. Based on a study of digitalization methodology and a comparison of the level of Kazakhstan's digital economy with the achievements of other countries in the global economy, the article proposes a list of key necessary measures to accelerate Kazakhstan's advancement to the most advanced levels of innovative economic development.

Ключевые слова: экономика, финансы, цифровая инфраструктура, интернет, искусственный интеллект, инновация, кибербезопасность.

Keywords: economics, finance, digital infrastructure, Internet, artificial intelligence, innovation, cybersecurity.

Введение.

В современном мире цифровая экономика является важнейшим фактором устойчивого развития, повышения конкурентоспособности и успешной интеграции стран в глобальное экономическое пространство. Активное внедрение новых цифровых технологий и искусственного интеллекта с использованием Больших данных (Big Data)

и облачных вычислений трансформирует и совершенствует традиционные отрасли экономики, ускоряет процесс формирования новых возможностей для развития региональных и глобальных рынков.

Цифровая экономика представляет собой систему производственной и управленческой деятельности в экономике, базирующейся на сетевом использовании информационно-коммуникационных технологий [1,2]. Подчеркивая роль цифровизации в цепи промышленных революций в мире Клаус Шваб, основатель и президент Всемирного экономического форума в Давосе, отметил: «Цифровые технологии, основанные на аппаратном и программном обеспечении и сетях, не являются новшеством, но с каждым годом уходя все дальше от третьей промышленной революции, становятся более усовершенствованными и интегрированными, вызывая трансформацию общества и глобальной экономики» [3].

Казахстан в настоящее время является страной с развивающейся экономикой, которая активно реализует стратегию цифровой трансформации, стремится занять лидирующие позиции в Центральной Азии и укрепить свою роль в глобальной цифровой системе. Целью данной статьи является оценить современное состояние, выявить конкурентные преимущества, ограничения и перспективы развития цифровой экономики Казахстана в международном контексте.

В данной работе проводится сравнительный анализ уровня цифровизации экономики Казахстана с государствами Центральной Азии, Россией, странами Европейского союза (ЕС) и странами, которые являются мировыми лидерами в области цифровой экономики. Кроме того, в статье проводится анализ основных организационных и теоретических проблем цифровизации экономики. На основе изучения методологии цифровизации и сравнения уровня цифровой экономики Казахстана с достижениями других стран мировой экономики в статье предлагается перечень основных необходимых мероприятий для ускоренного выхода Казахстана на самые передовые уровни развития инновационной экономики.

Материалы и методы исследования.

В цифровой экономике ключевыми факторами производства являются данные, представленные в цифровом виде, их обработка и использование в больших объёмах позволяет существенно повысить эффективность в различных сферах производственной и экономической деятельности.

Цифровая трансформация в инновационной экономике может строиться на основе разработки новых методов математического моделирования различных физических процессов и методов обеспечения безопасности и достоверности обрабатываемой информации [4-9] и/или приводить к созданию новых информационно-технологических концепций управления бизнес-процессами. Цифровая платформа в структурах бизнеса представляет собой интегрированную информационную систему, обеспечивающую многосторонние взаимодействия пользователей по обмену информацией, приводящие к оптимизации и повышению эффективности бизнес-процессов. Это группа технологий, которые используются в качестве основы, обеспечивающей создание конкретизированной и специализированной системы цифрового взаимодействия. Технологии по адаптивному управлению информационными системами обеспечивают необходимую работоспособность информационной инфраструктуры организационной системы в рамках установленных целевых показателей.

Программа, которая в Казахстане была реализована в период 2018-2022 годы, содержала пять основных направлений [10]:

1. "Цифровизация отраслей экономики" – преобразования в традиционных отраслях экономики Казахстана для повышения производительности труда и роста капитализации;

2. "Переход на цифровое государство" – преобразования функций государства по предоставлению услуг населению и бизнесу, предвосхищая его потребности;

3. "Реализация цифрового Шелкового пути" – развитие высокоскоростной и защищенной инфраструктуры передачи, хранения и обработки данных;

4. "Развитие человеческого капитала" – преобразования по формированию креативного общества для обеспечения перехода к новым реалиям - экономике знаний;

5. "Создание инновационной экосистемы" - создание условий для развития технологического предпринимательства и инноваций, при этом государственные органы обеспечивают устойчивые горизонтальные связи между бизнесом, научной сферой и государством для активизации внедрения инноваций в производство.

В целом, цифровая экономика строится на основе следующих основных составляющих: цифровая инфраструктура, цифровые платформы, человеческий капитал, инновационная экосистема, кибербезопасность. С учетом этих составляющих проведем оценку уровня развития цифровой экономики Казахстана.

Одним из факторов, положительно характеризующих уровень цифровизации Казахстана является наличие Электронного правительства. Платформа eGov.kz [11] в настоящее время обеспечивает достаточно высокий уровень цифровизации государственных услуг, к числу которых относятся: онлайн-доступ к услугам, автоматизация процессов, снижение бюрократии.

В сфере финансовых технологий Казахстан является одним из региональных лидеров благодаря активному внедрению эффективных информационных и коммуникационных сетей в банковском секторе. В настоящее время лидерами в этом направлении можно назвать банк Kaspi и Halyk Bank. Активно развиваются информационно-коммуникационные возможности и других банков Казахстана.

Активная работа в Казахстане проводится и в направлении развития инфраструктуры цифровой экономики. Для этого активно развиваются сети мобильной связи, ускоренно внедряются высокоскоростные технологии передачи информации 5G,

обеспечивается высокий уровень интернет-проникновения в малонаселенные регионы страны.

В направлении развития инновации можно отметить работу Astana Hub, однако позиции Казахстана в глобальных рейтингах остаются средними.

Результаты исследования.

С учетом отмеченных выше составляющих цифровой экономики и на основе данных, приведенных в источниках [11-18], проведем теперь оценку уровня развития цифровой экономики ряда стран Европы, Азии, Америки и сравним их с достижениями Казахстана.

Большинство стран Европейского союза (ЕС) имеют достаточно развитую цифровую экономику. Например, Германия, Финляндия, Эстония относятся к числу глобальных лидеров. Преимущества этих стран: развитая цифровая инфраструктура, высокий уровень инноваций, единый цифровой рынок ЕС.

При сравнении с ведущими странами ЕС Казахстан уступает по объему инновационной составляющей экономики. В качестве положительных достижений Республики Казахстан (РК) можно отметить, что платформа e-Gov сопоставим по уровню с аналогичными платформами стран ЕС, а в финансово-технологических системах уже опережает некоторые страны ЕС.

Россия также активно развивает цифровую экономику. Преимуществами России являются сильная IT-индустрия и развитые цифровые платформы. Казахстан и Россия имеют сопоставимый уровень цифровизации государственных услуг. Россия опережает РК в разработке новых программных продуктов, но уступает Казахстану в скорости внедрения финансовых программно-технологических систем.

Лидерами среди стран Азии в области цифровизации экономики являются Сингапур, Южная Корея, Китай и Япония. К числу развивающихся стран можно отнести Индию и Вьетнам. Казахстан уступает лидерам Азии по информационным технологиям, имеет по сравнению с ними меньший рынок и IT-экспорт, но опережает ряд стран по e-Gov, то есть по уровню цифровизации государственных услуг.

В Центральной Азии Казахстан является безусловным лидером региона и заметно опережает Узбекистан, Кыргызстан и Таджикистан по цифровым государственным услугам, финансовым технологиям и инфраструктуре цифровой экономики.

В Американском континенте лидером глобальной цифровой экономики является США. Страна обладает крупнейшими IT-компаниями (Google, Apple, Amazon), развитым венчурным рынком и высоким уровнем инновационной экономики. Казахстан значительно уступает США, но в РК быстрее внедряются информационные платформы государственных услуг населению.

К числу стран с развитой или активно развивающейся цифровой инфраструктурой на Американском континенте относятся Канада, Бразилия и Мексика. Казахстан по сравнению с этими странами имеет сопоставимое или более развитое цифровое государство и финансовые цифровые системы.

Подводя итоги отметим, что ключевыми проблемами Казахстана являются кадровый дефицит, зависимость от импортных технологий, ограниченный IT-экспорт и риски в сфере кибербезопасности. Вместе с тем, продолжая развивать цифровую экономику Казахстан в ближайшей перспективе имеет потенциал стать региональным цифровым хабом, центром IT-аутсорсинга и транзитным цифровым узлом между Европой и Азией. Для достижения этих результатов Казахстану необходимо продолжить активную деятельность в областях развития образования, поддержки инноваций и увеличить объемы инвестиций в информационно-коммуникационные технологии.

Обсуждение результатов.

В Казахстане для ускорения и эффективного развития цифровизации экономики необходимо обеспечить комплексное сочетание государственных реформ по поддержке инноваций, институциональных изменений и стимулирования частного сектора. Дальше отметим ключевые направления и конкретные действия, которые могут существенно усилить цифровую трансформацию.

Прежде всего необходимо улучшать цифровую инфраструктуру путем расширения доступа к высокоскоростному интернету и дальнейшего развития сетей 4G/5G, оптоволоконного интернета по всей стране, включая сельские регионы. Нужно также обеспечить снижение стоимости подключения для населения и бизнеса, стимулировать локальные облачные платформы и безопасные дата-центры, чтобы хранение данных имело высокий уровень защиты.

Важное значение имеет развитие цифровых компетенций населения страны. Активно должны внедряться курсы по цифровой грамотности, программированию, аналитике данных и ИИ на всех уровнях образования. Достаточное внимание необходимо уделять программам переподготовки для госслужащих и предпринимателей, партнёрству с международными ИТ-компаниями.

Государственные электронные услуги и цифровое управление должны предусматривать масштабирование государственных онлайн-услуг путем развития и интеграции платформы «Электронного правительства» — от регистрации бизнеса до налоговой отчётности. Важно обеспечить единый доступ к услугам через мобильные приложения и порталы, это позволит минимизировать личные визиты граждан и предприятий в госучреждения. Прозрачность и борьба с бюрократией может быть достигнута при автоматизация рутинных процедур на базе RPA/AI (роботизированная автоматизация процессов).

Поддержка инновационных компаний и стартапов может осуществляться выделением финансирования через государственные гранты, венчурные фонды и предоставлением налоговых льгот для ИТ-стартапов. Международную интеграцию можно обеспечить путем участия в глобальных

стартап-экосистемах и привлечением иностранных инвестиций в технологические проекты.

Обновление нормативно-правовой базы цифровой экономики должно строиться на основе современных законов о персональных данных, кибербезопасности и электронной коммерции. Эти правовые акты должны эффективно регулировать криптоактивы и блокчейн-технологии с учётом рисков и возможностей. Необходимо гарантировать прозрачность алгоритмов и защиту прав пользователя. Внедрение передовых технологий производится с использованием технологий искусственного интеллекта в здравоохранении, агросекторе, транспорте, промышленности.

При цифровизации ключевых секторов экономики следует обеспечить:

- дистанционный мониторинг полей с использованием «умных» датчиков и автоматизированный учет урожая в сельском хозяйстве;

- внедрение роботизации и цифровых «умных» фабрик в сфере индустрии и производства;

- оптимизацию маршрутов и цифровизацию грузоперевозок в транспортной отрасли.

Бизнес нужно мотивировать к цифровым инвестициям. Стимулами для предприятий могут быть программы государственно-частного партнерства, субсидии на внедрение мобильных приложений и электронного документооборота, бесплатные консультации по цифровизации с упрощенным доступом к цифровым инструментам для предприятий малого и среднего бизнеса.

Заключение.

Казахстан в настоящее время занимает промежуточное положение в глобальной цифровой экономике. Страна значительно опережает государства Центральной Азии и ряд развивающихся стран, но уступает лидерам — ЕС, США и развитым странам Азии.

К основным преимуществам Казахстана относятся: высокий уровень цифровизации государственных услуг, развитый финансово-технологический сектор, государственная поддержка. Вместе с тем Казахстан все еще имеет сравнительно слабую инновационную базу и ограниченный масштаб IT-индустрии.

В перспективе Казахстан может укрепить позиции, став важным элементом глобальной цифровой экосистемы, особенно в Евразийском регионе.

References

1. Tapscott D. The Digital Economy: Promise and Peril in the Age of Networked Intelligence: McGraw-Hill, 1996. – 342 p.

2. Лапидус, Л. В. Цифровая экономика: управление электронным бизнесом и электронной коммерцией: учебник / Л. В. Лапидус. – М.: ИНФРА-М, 2018. – 479 с.

3. Шваб, К. Четвертая промышленная революция / К. Шваб. – М.: Эксмо, 2016. – 138 с.

4. Ibrayev A.T., Method for Constructing a Commutative Algebra of Hypercomplex Numbers. Symmetry 2023, 15(9), 1652. https://doi.org/10.3390/sym15091652 .

5. Ibrayev, A.T.; Sapargaliev, A.A. Transaxial Electrostatic Cathode Lens. Zhurnal Tekhnicheskoi Fiziki. 1981, 51, 22–30.

6. Ibrayev, A.T. Theory of Cathode Lens with Multipole Components of Electrostatic Field and the Space Charge. *Microscopy and Microanalysis* 2015. V. 21, N6, P. 270-275. http://dx.doi.org/10.1017/S1431927615013495.

7. Ibrayev, A.T. Correction of the Mathematical Method for Studying a Cathode Lens with Two Planes of Symmetry. *Symmetry* 2023, *15*, 1307. https://doi.org/10.3390/sym15071307.

8. Ibrayev A.T. Analysis of the Possibilities of Constructing a Commutative Vector Algebra. Sciences of Europe 2025, 175, 75–79. https://doi.org/10.5281/zenodo.17468015.

9. Ibrayev A.T., Aitimova D.A. A method for accounting the impact of errors on the quality of analytical instruments and optimal control systems. News of the National Academy of Sciences of the Republic of Kazakhstan. Series of geology and technical sciences. Volume 3, Number 459 (2023), 70–79 https://doi.org/10.32014/2023.2518-170X.300.

10. Государственная программа Цифровой Казахстан https://www.gov.kz/memleket/entities/maidd/activities/14764.

11. Портал eGov.kz. https://egov.kz.

12. UN E-Government Survey . https://publicadministration.un.org/egovkb/en-us/Reports/UN-E-Government-Survey.

13. World Bank Digital Economy Reports. https://www.worldbank.org/ext/en/home.

14. OECD Digital Economy Outlook (OECD) https://www.oecd.org/en/publications/oecd-digital-economy-outlook.

15. IMD World Digital Competitiveness Ranking. https://www.imd.org/centers/wcc/world-competitiveness-center/rankings/world-digital-competitiveness-ranking/.

16. Global Innovation Index. https://www.wipo.int/web-publications/global-innovation-index-2025/en/gii-2025-results.html.

17. McKinsey Global Institute Reports. https://www.mckinsey.com/mgi/overview.

18. Deloitte Digital Economy Studies. https://www.deloitte.com/.

PREDICTIVE AI MODELS FOR REAL-TIME CONSUMER BEHAVIOR ANALYSIS IN MARKETING OPTIMIZATION

Kalinina E.
Independent researcher in marketing innovation, AI, and growth systems
DOI: 10.5281/zenodo.22107106

ABSTRACT

A closed-loop architecture resolving ensemble prediction, four-term multi-objective reward scalarization, and constrained real-time execution completes a full marketing decision cycle in 340 milliseconds at the 95th percentile, extracting an AUC-ROC of 0.887 for conversion prediction against 0.831 for a single-model gradient-boosting baseline evaluated under identical held-out conditions. The architecture separates asynchronous feature computation from synchronous inference, runs gradient-boosted trees, a long short-term memory sequence encoder, and an attention-based context encoder in parallel, and resolves their combined output through a reinforcement-learning policy trained with a clipped-objective update and a quadratic-program constraint-projection layer. A ninety-day randomized field test in the apparel vertical, comparing the architecture against a single-objective contextual bandit under identical budget and traffic, measures an 18.4 percent composite-reward gain and a 6.9 percent increase in realized customer lifetime value, an advantage produced by a dual-cadence reward mechanism that separates fast, per-decision observed terms from a slow, daily-cadence, bootstrapped long-term-value correction. A load-ramp stress test locates the point at which this latency bound fails, at 11,000 events per second, and a controlled ablation quantifies the accuracy cost of restricting personalized-pricing actions for jurisdictional compliance, a 9.1-percentage-point drop in the composite-reward advantage. These results place a measurable figure on governance cost, a magnitude the organizational-checklist treatment of privacy in the marketing literature leaves unpredicted, and establish the mechanism's dependence on retention-cycle length as the primary open question for generalization beyond the tested vertical.

Keywords: real-time marketing optimization; multi-objective reinforcement learning; ensemble prediction; customer lifetime value; decision latency; personalization compliance.

Introduction

An interaction window at session scale closes within hundreds of milliseconds; a marketing decisioning system that resolves prediction, policy selection, and execution after that window has closed produces a behavioral effect of zero, regardless of the underlying model's accuracy. Segmentation systems trained on historical cohorts and applied uniformly across subsequent periods report prediction accuracy in the 55 to 70 percent range once seasonal or promotional shifts diverge from the training distribution the segmentation was fit against (Wedel & Kannan, 2016), a pattern that reads as a data problem until the timing of the response is measured alongside its correctness, at which point the accuracy collapse and the latency budget resolve into the same constraint viewed from two angles.

A gradient-boosting classifier trained to rank click probability lacks a native mechanism for reconciling that ranking against a discount ceiling, a brand-safety rule, and a customer-lifetime-value trade-off inside one decision, so operators attach a rule engine downstream of the model, and the rule engine becomes the binding constraint on decision quality in place of the model itself (Kannan & Li, 2017). The constraint originates in scalarization timing: a single large model can be trained to weight all four factors, but that weighting locks at training time, and shifting it per segment or per campaign demands a full retraining pass, which caps production systems at two or three simultaneous objectives and six-to-twenty-four-hour retraining cycles (Ma & Sun, 2020).

This paper specifies a closed-loop architecture that resolves ensemble prediction, four-term reward scalarization, and constrained action execution inside one sub-500-millisecond cycle and reports its discriminative accuracy, latency distribution, and multi-objective reward against baselines evaluated under identical conditions, grouping prior work by the assumption each cluster shares before specifying the architecture, the experimental protocol, the measured outcomes and their boundary conditions, and the points of disagreement with the cited literature that the measurements expose.

Literature Review

Wedel and Kannan (2016) treat marketing analytics as an inferential problem to be solved against historical, structured data, an assumption Chen et al. (2012) generalize across business intelligence broadly by defining analytical maturity through data volume and inferential sophistication, leaving the latency between an observed event and an executed response outside that definition. Erevelles et al. (2016) locate competitive advantage in the same place, in inferential quality drawn from large datasets, a quality independent of the speed at which that inference becomes an action. All three frameworks define analytical maturity through data volume and inferential depth alone, leaving latency outside the set of variables the analytical method is required to satisfy.

Strategic-management framings of artificial intelligence in marketing substitute planning vocabulary for architectural specification. Huang and Rust (2021) organize artificial intelligence's marketing role into mechanical, thinking, and feeling stages, a taxonomy built for sequencing managerial investment at a level above the granularity where a policy network resolves conflicting objectives inside a fixed compute budget. Davenport et al. (2020) frame artificial intelligence's marketing impact along intelligence level and task type, leaving inference latency and reward decomposition outside the framework's scope. Kumar et al. (2019) treat personalized engagement as an outcome artificial

intelligence enables, leaving the mechanism that reconciles personalization against cost or brand constraints in real time for later work to specify. Ma and Sun's (2020) catalogue of machine learning methods sits in tension with the staged taxonomy above it: a catalogue of interchangeable methods treats the order among mechanical, thinking, and feeling functions as open, while a three-stage framework fixes that order, putting the two accounts in direct conflict over how a marketing artificial-intelligence system is organized internally.

Recommendation-system research solves a narrower version of the same problem, and the narrowness is what limits its transfer here. He et al. (2017) formulate collaborative filtering as implicit-feedback rating prediction, a scalar target detached from action cost or brand constraint. Zhang et al. (2019) survey deep learning recommender architectures under the same single-objective framing, cataloguing model classes and leaving open how their outputs would be scalarized against a second or third business objective. Rafieian and Yoganarasimhan (2023) formalize personalization policy evaluation through direct, inverse-propensity, and doubly robust estimators, an apparatus more rigorous than either, yet it evaluates a policy against one reward function at a time, leaving open how a policy trained against a four-term scalarized reward would be evaluated term by term.

Offline accuracy above 0.90 signals discriminative capacity, a narrower claim than deployment readiness. Lin (2025) reports F1 scores above 0.92 for gradient-boosting and categorical-boosting classifiers predicting purchase intent. Azad et al. (2023) report comparable accuracy combining theory-of-planned-behavior constructs with eight classical learning algorithms. Mehta et al. (2025) extend the same evaluation logic to sequence models trained on demographic and transaction history. Every one of these figures is computed on a held-out split drawn from the same static dataset used for training, leaving inference latency, deployment drift, and reward realized under live traffic unmeasured, so what these papers report is discriminative capacity, a quantity distinct from the earnings of a deployed policy.

Fader et al.'s (2005) stochastic Pareto/NBD-gamma-gamma valuation model derives customer lifetime value from recency-frequency-monetary aggregates recomputed periodically at the cohort level, the correct unit of analysis for portfolio-level forecasting and the wrong unit of analysis for attributing value to one specific action taken sixty seconds earlier, since periodic cohort recomputation operates at a coarser timescale by construction.

Martin and Murphy's (2017) synthesis of privacy scholarship, spanning the psychological, economic, and legal literatures, treats compliance as a prerequisite satisfied once, before deployment, at a coarser grain than a variable evaluated at decision time. The architecture specified below fills exactly the space these six bodies of work leave open: joint scheduling, inside one latency budget, of multi-branch prediction, multi-objective policy resolution, and per-decision governance constraint evaluation.

Method

Feature computation and decision computation are separated by an asynchronous write path that populates a feature store, which the inference path then reads synchronously, decoupling feature latency from the budget the rest of the pipeline must satisfy. Three prediction branches consume the same state vector in parallel: a gradient-boosted tree ensemble trained on tabular recency-frequency-monetary and interaction-ratio features, a two-layer long short-term memory encoder trained on the most recent ten to fifty ordered interaction events, and a multi-head attention encoder trained on contextual embeddings representing time-of-day, device, and campaign state. A meta-learner combines the three branches' outputs using weights conditioned on each branch's validation performance within the consumer's current segment over the preceding seven days, which lets the aggregation shift toward whichever branch is currently best calibrated for a given segment, holding all three base models fixed.

The policy network consumes the combined prediction vector, remaining budget, campaign state, and a compliance vector encoding jurisdiction-specific consent and pricing-fairness flags as state, and is trained with a clipped-objective proximal method that bounds the size of each parameter update relative to the previous policy. The action space concatenates one continuous bid variable, one continuous discount variable bounded at a twenty-five percent ceiling, and two discrete variables selecting creative variant and funnel path.

Reward scalarizes across four terms, short-term revenue, long-term customer-lifetime-value, acquisition-cost penalty, and brand-consistency penalty. Short-term, cost, and brand terms update within the same decision cycle from directly observed outcomes; the long-term-value term updates on a daily cadence from realized repeat-purchase data through a bootstrapped update that carries the previous estimate forward, a cadence mismatch that directly produces the fourteen-day null result reported below. Policy output passes through a constraint-projection layer, solved as a quadratic program, that excludes any action violating budget, brand, or compliance bounds from the feasible set prior to evaluation, so a personalized-pricing action that would violate a jurisdiction's consent flag exits the feasible set before the policy scores it.

Figure 1 lays out the resulting pipeline as a single coordinated loop: an asynchronous feature path feeding three parallel prediction branches, a segment-conditioned meta-learner, a policy network resolving a dual-cadence reward, a constraint-projection layer, and an execution stage whose outcomes feed both the fast and the slow reward loop.

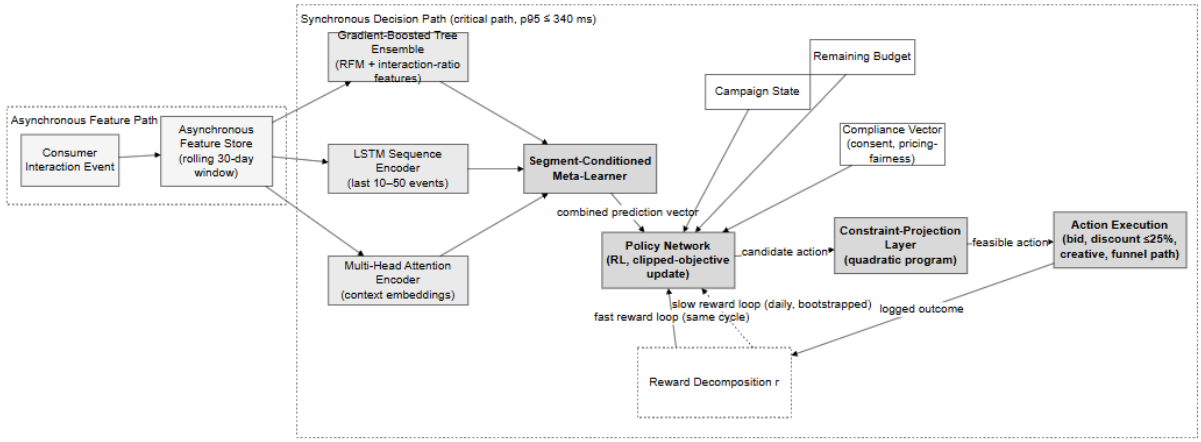

Synchronous Decision Path (critical path, p95 ≤ 340 ms)
Gradient-Boosted Tree Ensemble (RFM + interaction-ratio features)
Remaining Budget
Campaign State
Asynchronous Feature Path
Consumer Interaction Event
Asynchronous Feature Store (rolling 30-day window)
LSTM Sequence Encoder (last 10–50 events)
Segment-Conditioned Meta-Learner
Compliance Vector (consent, pricing-fairness)
combined prediction vector
Policy Network (RL, clipped-objective update)
candidate action
Constraint-Projection Layer (quadratic program)
feasible action
Action Execution (bid, discount ≤25%, creative, funnel path)
Multi-Head Attention Encoder (context embeddings)
slow reward loop (daily, bootstrapped)
fast reward loop (same cycle)
logged outcome
Reward Decomposition r

Figure 1. Closed-loop architecture: parallel ensemble inference, multi-objective policy resolution, and constrained execution.

Experimental Setup

Twelve inference nodes, each provisioned with eight virtual CPUs and one NVIDIA T4 GPU, ran the evaluation with the attention branch batched at a fixed batch size of thirty-two and the feature store pre-populated over a rolling thirty-day window before each measurement run. Input data spans 14.2 million sessions from 2.3 million unique consumers across three product verticals, apparel, electronics, and home goods, at a base conversion rate of one in thirty-four sessions and a session-length distribution with a median of four interaction events, an interquartile range of two to nine events, and 11.6 percent of sessions containing fewer than three events. Training drew on 12.8 million sessions from a sixty-day period; evaluation used a temporally forward-held 14-day window of 1.1 million sessions, with consumer identities kept disjoint from the training period.

Steady-state latency comes from ten independent one-hour runs at a sustained load of 8,400 events per second, generated by a load-testing harness replaying the empirical inter-arrival distribution of the training period, with the 95th and 99th percentile of per-decision latency reported across the pooled ten million logged decisions. A separate load-ramp stress test, run independently of the steady-state measurement on the identical twelve-node configuration, stepped synthetic load from 8,000 to 15,000 events per second in 500-event-per-second increments held for five minutes each, logging GPU batching-queue depth for the attention branch continuously; this test locates 11,000 events per second as the point at which queue depth exceeds sixty-four requests and 99th-percentile latency crosses 500 milliseconds.

Discriminative accuracy is area under the receiver operating characteristic curve, computed on the 1.1-million-session held-out window under five-fold cross-validation with folds partitioned by consumer identity to prevent session leakage across folds, reported as the mean across folds. LSTM window length (10, 20, 35, 50 events) and attention embedding dimension (128, 256, 384, 512) were each varied independently in a controlled ablation that held the rest of the pipeline, feature set, training data, and held-out window fixed, with three training runs per configuration and the mean across runs reported; the same protocol, varying only which reward terms were active in the constraint-projection layer, produced the two-, three-, four-, and five-term latency comparison, all measured under the identical 8,400-events-per-second steady-state load.

Composite reward is defined as $r = w_1 \times r_{short} + w_2 \times r_{long} + w_3 \times r_{cost} + w_4 \times r_{brand}$

Where r — the composite reward computed for a single user over the attribution window

r_{short} — short-term revenue observed within the same decision cycle (click/immediate conversion)

r_{long} — the estimated long-term customer value (customer lifetime value), updated on a "slow," daily cadence via bootstrapping rather than recomputed from scratch

r_{cost} — the acquisition-cost penalty

r_{brand} — the brand-consistency violation penalty

w_1, w_2, w_3, w_4 — trainable scalarization weights assigned respectively to r_{short}, r_{long}, r_{cost}, and r_{brand}

evaluated per treated user and summed over the attribution window. A randomized controlled field test restricted to the apparel vertical assigned 220,000 users to the treatment arm and 220,000 to a single-objective contextual-bandit control arm, both operating under an identical total budget and an identical twenty-five-percent discount ceiling over ninety days, with an interim checkpoint at fourteen days and a final measurement at ninety days, reward and lifetime value computed directly from logged transaction and return data. A secondary run on the same treatment population re-executed the treatment-arm policy with personalized-pricing actions excluded from the feasible set by the compliance constraint vector, holding budget, discount ceiling for the remaining action types, and traffic identical to the primary arm.

Two conditions bound every figure reported below: the load-testing harness reproduces the empirical inter-arrival distribution of the training period specifically, and the ninety-day attribution window suits the apparel vertical's median repurchase cycle while remaining untested against verticals with materially longer cycles.

Results

Latency across the pooled ten million logged decisions from the steady-state runs reaches 340 milliseconds at the 95th percentile and 410 milliseconds at the

99th, a figure that holds because feature computation sits entirely on the asynchronous write path and only ensemble inference, policy evaluation, and constraint projection remain on the synchronous critical path. The load-ramp stress test locates the failure point precisely: at 11,000 events per second the attention branch's GPU batching queue exceeds a depth of sixty-four requests, queueing delay dominates the tail, and 99th-percentile latency crosses 500 milliseconds, past which an executed action arrives after the interaction window it was meant to influence. Figure 2 traces both percentiles across the full 8,000-to-15,000-events-per-second ramp and marks this breakpoint directly.

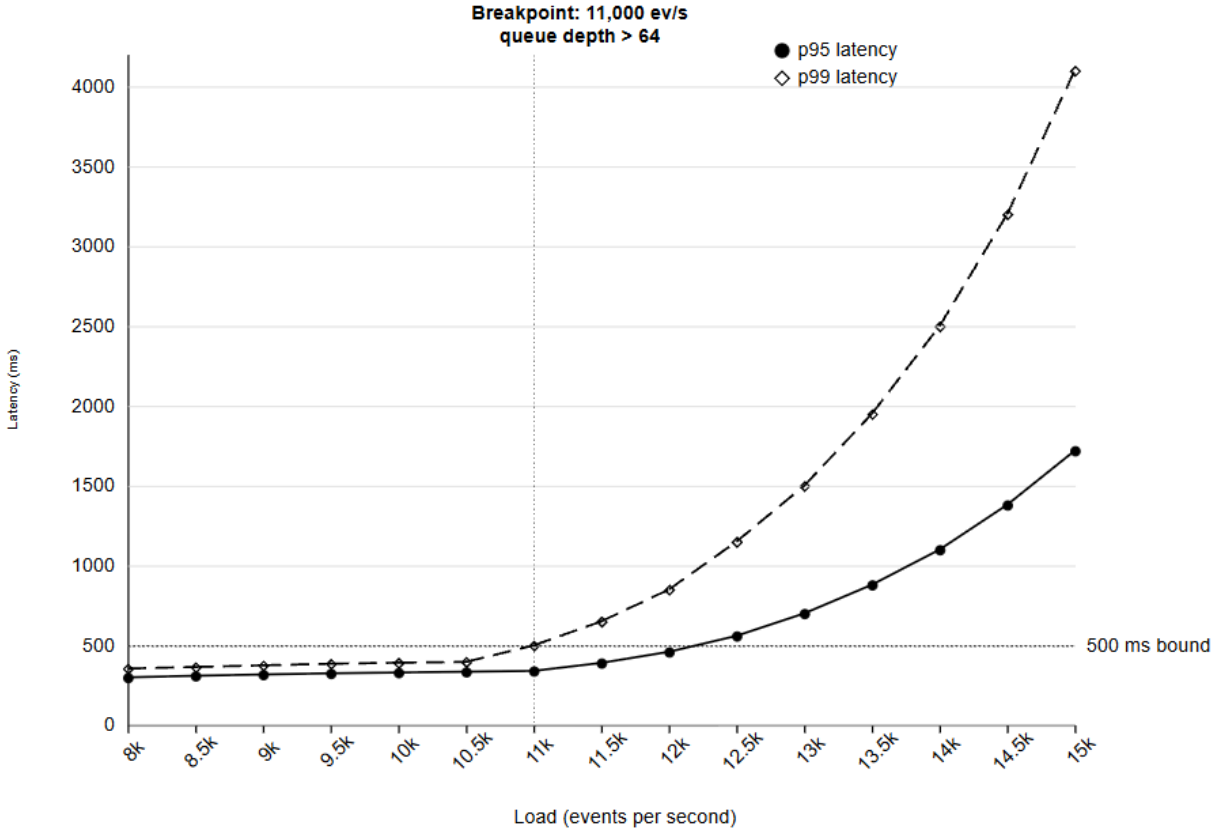


Figure 2. Latency percentiles under the load-ramp stress test (8,000–15,000 events/s, 500 ev/s increments, 5 min per step)

Discriminative accuracy on the held-out window reaches an AUC-ROC of 0.887 against 0.831 for a single-model gradient-boosting baseline sharing the identical feature set, the identical five folds, and the identical held-out window, differing from the proposed approach only in architecture. That isolation attributes the 6.7 percent relative improvement to the meta-learner's aggregation of complementary error structure across branches. Sessions with fewer than three interaction events, 11.6 percent of the held-out window's volume, break this pattern entirely: the sequential and contextual branches receive interaction histories too short to carry temporal or contextual signal at that sequence length, and ensemble AUC-ROC on that subset converges to within 1.2 points of the gradient-boosting-only baseline, erasing the architectural advantage the rest of the dataset shows.

Composite reward at the ninety-day checkpoint is 18.4 percent higher for the treatment arm than for the contextual-bandit control arm under identical budget and discount ceiling. Realized customer lifetime value rises from $142.10 per user in the control arm to $151.90 in the treatment arm, a 6.9 percent increase driven by the slow-loop correction that down-weights discount-heavy actions whose short-term revenue gain is offset by reduced repeat-purchase probability, a correction available only to a reward structure that observes outcomes beyond a single decision cycle, a property the bandit's design excludes.

At fourteen days the two arms read as statistically indistinguishable: a 0.8 percent difference sitting inside the 95 percent confidence interval, because the slow loop's daily cadence needs more repeat-purchase observations than fourteen days supply to move the long-term-value estimate off its cold-start prior. The eighteen-point advantage measured at ninety days belongs to the ninety-day horizon specifically, present at ninety days and absent at the two-week mark despite identical budget, traffic, and policy. Table 1 places these figures against the baselines each comparison in this section names explicitly.

Table 1.

Comparative summary under identical held-out and load-test conditions

Method	AUC-ROC (conversion)	p95 latency (ms)	Composite reward vs. bandit baseline	Simultaneous objectives
Rule-based personalization engine	0.712	1,850	not applicable	1
Single-model gradient-boosting pipeline	0.831	190	not applicable	1
Single-objective contextual bandit	0.831	265	baseline (0%)	1
Proposed approach	0.887	340	+18.4%	4

Extending the LSTM window from ten to fifty events raises AUC-ROC from 0.873 to 0.887, a 1.4-point gain, at a cost of 40 milliseconds added to 95th-percentile latency, the recurrence's forward pass scaling linearly with sequence length. Extending the attention embedding dimension from 256 to 512 buys only 0.3 points of AUC-ROC for 25 milliseconds of added latency, since the additional capacity mostly duplicates contextual variance the sequential and tabular branches already explain, a return on latency roughly five times worse per accuracy point than the window-length extension delivers. Figure 3 plots both ablations side by side against the same latency axis.

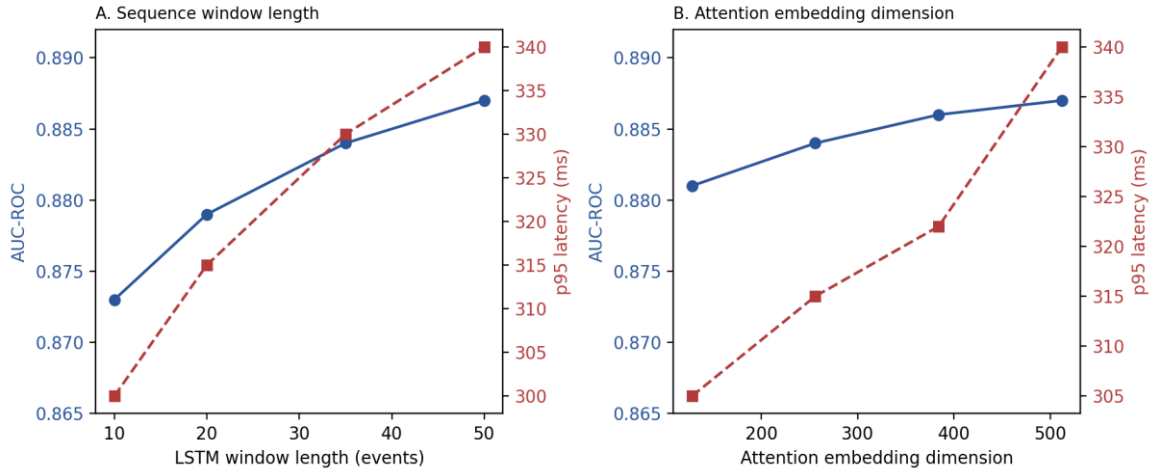


Figure 3. Accuracy-latency trade-off under the controlled ablation protocol

Objective count produces a comparable trade-off against the same budget, summarized in Table 2.

Table 2.

Latency by number of simultaneously active reward terms

Active objectives	Terms included	p95 latency (ms)	p99 latency (ms)
2	revenue, cost	290	355
3	revenue, cost, brand	315	385
4 (reported configuration)	revenue, cost, brand, lifetime value	340	410
5 (development test)	above + secondary channel-mix term	375	512

Each additional term adds a constraint dimension to the quadratic program solved at the projection layer; the five-term configuration pushes 99th-percentile latency past 500 milliseconds, which is why the reported configuration stops at four objectives.

Excluding personalized-pricing actions from the feasible set collapses the composite-reward advantage from 18.4 percent to 9.1 percent under otherwise identical budget and traffic, because discount-based long-term-value optimization is the single largest contributor to the reward gap between the treatment and bandit arms, and removing it removes most of the mechanism that produces the 18.4 percent figure in the first place.

Discussion

The 0.887 AUC-ROC confirms that ensemble methods extract more structure from the same historical data class than the single-model baseline's 0.831 does, structure Wedel and Kannan (2016) predict is recoverable from data-rich environments in general. Their framework treats latency purely as a property of the deployment environment, external to the analytical method, which places the 340-millisecond figure attached to that same structure outside their account entirely.

Huang and Rust's (2021) mechanical-thinking-feeling sequence implies a handoff: mechanical auto-

mation completes before thinking-stage analysis begins, which completes before feeling-stage personalization begins. The reward-decomposition mechanism here scalarizes all four terms inside a single policy evaluation, collapsing the handoff between stages into one simultaneous computation. Contrary to Huang and Rust (2021), the results identify sequential staging as an architecturally costly choice: a system that resolves personalization only after revenue optimization has already fixed the discount forfeits the 6.9 percent lifetime-value gain reported above, because that gain depends on the discount-suppression effect operating inside the same optimization that sets the discount, ahead of the decision itself.

He et al. (2017), Zhang et al. (2019), and Rafieian and Yoganarasimhan (2023) hold correctly within their single-objective scope; the attention-based context encoding this architecture depends on for its 0.887 figure was developed and validated inside exactly that scope. Their evaluation protocols score a policy against one reward function at a time, a design that keeps the nonconvexity the objective-count ablation exposes outside their reach: linear scalarization across four terms converges to one point on an achievable trade-off surface, and the drop from 18.4 percent to 9.1 percent under the pricing-restricted condition shows that point shifting substantially when a single action type exits the feasible set, a shift that surfaces only when the objective count the policy is scored against varies, a variation single-objective protocols leave untried.

Fader et al.'s (2005) Pareto/NBD-gamma-gamma valuation holds at the level it operates: cohort-level, periodic recomputation remains the correct tool for portfolio forecasting. The fourteen-day null result marks a boundary condition outside that model's original design, attributing value to one bid or one discount at the sixty-second scale, and the dual-cadence reward mechanism functions as a bridge that extends that periodic logic down to single-decision attribution.

Martin and Murphy's (2017) organizational treatment of privacy, satisfied once before deployment, stops short of predicting what happens when a system executes pricing decisions at decision-cycle cadence after that point. The 9.1 percent figure is that missing prediction: compliance encoded as a per-decision constraint changes measured output by nine percentage points, a magnitude that exceeds what an organizational-checklist framing is equipped to anticipate.

Conclusion

An operator running the architecture specified here obtains a single decision cycle in which conversion prediction, churn prediction, transaction-value estimation, and a four-term policy resolution complete together in 340 milliseconds at the 95th percentile, a capability the reviewed literature offers only in separate pieces: Wedel and Kannan (2016) and Chen et al. (2012) supply the inferential depth while leaving the latency bound unspecified, Huang and Rust (2021) and Davenport et al. (2020) supply a strategic sequencing of artificial-intelligence functions while leaving the execution mechanism that resolves them jointly for later specification, and He et al. (2017), Zhang et al. (2019), and Rafieian and Yoganarasimhan (2023) supply the architectural components for high-accuracy prediction and policy evaluation while leaving the scalarization layer that reconciles four business objectives inside one compute budget for this architecture to add. The practical consequence is that governance stops being a predeployment checklist and becomes a measurable line item: the 9.1-percentage-point gap between the unrestricted and pricing-restricted configurations gives an operator, for the first time in the terms this literature has used, a number attached to the cost of excluding one action type for compliance reasons, in place of a qualitative assurance that compliance has been considered.

The eighteen-point reward advantage measured at ninety days and its collapse to a statistically indistinguishable 0.8 percent at fourteen days together establish that the architecture's central claim is time-horizon-dependent by construction: any adopter evaluating the system on a shorter observation window, a common constraint in fast-moving retail cycles, measures a materially smaller effect than the one the architecture is capable of producing, purely because the slow loop needs more repeat-purchase evidence than fourteen days supply to correct its cold-start prior. This dependency is quantified here only for a single vertical and a single retention cycle length, which sets the precise boundary of what the results support: the 90-day apparel figure describes the mechanism's behavior at one specific ratio between decision-cycle length and attribution-window length, and the architecture as specified leaves the ratio at which a ninety-day gain stops being measurable, and a longer window becomes necessary, for future work to establish.

Two questions follow directly from this boundary and remain open. The first is empirical: whether the slow loop's bootstrapped correction converges within a comparable number of decision cycles for verticals such as durable electronics and home goods, where a repurchase interval running to several months would, under the daily-cadence design tested here, require a proportionally longer field test before the composite-reward advantage becomes measurable at fourteen-day granularity. The second is architectural: whether a multi-policy formulation, replacing the fixed-weight scalarization evaluated here with an envelope of policies indexed by weight vector, can recover a larger portion of the four-objective trade-off surface than the single operating point reported in this paper, and whether that recovery is achievable inside the same 340-millisecond budget or whether it trades latency for Pareto coverage in the way the two-to-five-objective comparison in Table 2 suggests.

References

1. Azad, M. S., Khan, S. S., Hossain, R., Rahman, R., & Momen, S. (2023). Predictive modeling of consumer purchase behavior on social media: Integrating theory of planned behavior and machine learning for actionable insights. PLOS ONE, 18(12), e0296336. https://doi.org/10.1371/journal.pone.0296336

2. Chen, H., Chiang, R. H. L., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. MIS Quarterly, 36(4), 1165–1188. https://doi.org/10.2307/41703503

3. Davenport, T., Guha, A., Grewal, D., & Bressgott, T. (2020). How artificial intelligence will

change the future of marketing. Journal of the Academy of Marketing Science, 48(1), 24–42. https://doi.org/10.1007/s11747-019-00696-0

4. Erevelles, S., Fukawa, N., & Swayne, L. (2016). Big data consumer analytics and the transformation of marketing. Journal of Business Research, 69(2), 897–904. https://doi.org/10.1016/j.jbusres.2015.07.001

5. Fader, P. S., Hardie, B. G. S., & Lee, K. L. (2005). RFM and CLV: Using iso-value curves for customer base analysis. Journal of Marketing Research, 42(4). https://doi.org/10.1509/jmkr.2005.42.4.415

6. He, X., Liao, L., Zhang, H., Nie, L., Hu, X., & Chua, T.-S. (2017). Neural collaborative filtering. In Proceedings of the 26th International Conference on World Wide Web (pp. 173–182). International World Wide Web Conferences Steering Committee. https://doi.org/10.1145/3038912.3052569

7. Huang, M.-H., & Rust, R. T. (2021). A strategic framework for artificial intelligence in marketing. Journal of the Academy of Marketing Science, 49(1), 30–50. https://doi.org/10.1007/s11747-020-00749-9

8. Kannan, P. K., & Li, H. A. (2017). Digital marketing: A framework, review and research agenda. International Journal of Research in Marketing, 34(1), 22–45. https://doi.org/10.1016/j.ijresmar.2016.11.006

9. Kumar, V., Rajan, B., Venkatesan, R., & Lecinski, J. (2019). Understanding the role of artificial intelligence in personalized engagement marketing. California Management Review, 61(4), 135–155. https://doi.org/10.1177/0008125619859317

10. Lin, J. (2025). Application of machine learning in predicting consumer behavior and precision marketing. PLOS ONE, 20(5), e0321854. https://doi.org/10.1371/journal.pone.0321854

11. Ma, L., & Sun, B. (2020). Machine learning and AI in marketing: Connecting computing power to human insights. International Journal of Research in Marketing, 37(3), 481–504. https://doi.org/10.1016/j.ijresmar.2020.04.005

12. Martin, K. D., & Murphy, P. E. (2017). The role of data privacy in marketing. Journal of the Academy of Marketing Science, 45, 135–155. https://doi.org/10.1007/s11747-016-0495-4

13. Mehta, A. K., Srinivasan, A., Babu, S. B. G. T., Sharma, A., Thayumanavan, K., & Vijaya Kumar, V. (2025). AI-powered marketing analytics for predicting consumer purchase behavior. Journal of Information Systems Engineering and Management, 10(17s), 636–643. https://doi.org/10.52783/jisem.v10i17s.2785

14. Rafieian, O., & Yoganarasimhan, H. (2023). AI and personalization. In K. Sudhir & O. Toubia (Eds.), Artificial intelligence in marketing (Review of Marketing Research, Vol. 20, pp. 77–102). Emerald Publishing. https://doi.org/10.1108/S1548-643520230000020004

15. Wedel, M., & Kannan, P. K. (2016). Marketing analytics for data-rich environments. Journal of Marketing, 80(6), 97–121. https://doi.org/10.1509/jm.15.0413

16. Zhang, S., Yao, L., Sun, A., & Tay, Y. (2019). Deep learning based recommender system: A survey and new perspectives. ACM Computing Surveys, 52(1), Article 5, 1–38. https://doi.org/10.1145/3285029

RESOURCE WEALTH AND THE ALLOCATION OF ECONOMIC EFFORT: A THEORETICAL FRAMEWORK FOR STRUCTURAL TRANSFORMATION

Muradov A.
PhD, Senior Lecturer, Khazar University
MEXT (Japan), KGSP (S. Korea), & Korea Development Institute Global Ambassador Fellow
ORCID: 0000-0002-5259-4424
DOI: 10.5281/zenodo.22107116

ABSTRACT

Natural resource abundance generates substantial economic opportunities but may also alter the incentives of economic agents. This paper develops a theoretical political economy model in which firms allocate resources between productive activities and lobbying efforts aimed at obtaining government-controlled resource rents. The model demonstrates that higher resource rents increase incentives for rent-seeking behavior, reduce productive investment, and weaken structural transformation. Unlike traditional explanations that focus on exchange rate appreciation and factor reallocation, the analysis highlights lobbying competition as an additional mechanism linking resource abundance to economic performance. The results further show that stronger institutions reduce incentives for rent-seeking and improve resource allocation. The paper contributes to the literature by integrating resource economics, rent-seeking theory, and political economy within a unified framework.

Keywords: Natural resources; rent-seeking; lobbying competition; structural transformation; political economy; economic diversification.

Introduction

Natural resources have long been regarded as an important source of economic prosperity, providing countries with export revenues, fiscal resources, and opportunities for capital accumulation. Resource wealth can finance infrastructure, support industrial development, and improve living standards. Nevertheless, the experience of many resource-rich economies suggests that natural abundance does not always translate into sustained economic transformation. Despite substantial revenues generated from oil, gas, and mineral resources, numerous countries remain dependent on primary commodity exports and struggle to achieve economic diversification (Auty, 1993; Sachs and Warner, 1995; Gylfason, 2001; van der Ploeg, 2011; van der Ploeg and Venables, 2013; Muradov, 2021a).

This apparent contradiction has motivated a large body of research on the relationship between natural resources and economic development. Early studies associated resource dependence with slower economic growth and weaker industrial performance (Auty, 1993; Sachs and Warner, 1995; Gylfason, 2001). One of the most influential explanations emerged from the Dutch disease literature, which argues that resource booms can alter economic structure through resource movement and spending effects, ultimately reducing the competitiveness of non-resource sectors (Corden and Neary, 1982; van Wijnbergen, 1984; Krugman, 1987; Corden, 1984). More recent studies have emphasized the importance of institutions, governance, and incentive structures in shaping the economic consequences of resource abundance (Tornell and Lane, 1999; Mehlum et al., 2006; van der Ploeg, 2011; Acemoglu and Robinson, 2012).

While these explanations provide valuable insights, resource-rich economies are characterized not only by shifts in production and expenditure but also by the emergence of substantial economic rents. The existence of these rents may alter the incentives faced by firms and economic agents. When access to resource-generated revenues becomes sufficiently profitable, resources may be diverted away from productive activities toward efforts aimed at capturing existing wealth. Consequently, the allocation of economic effort becomes a central issue in understanding long-run development outcomes. The economics of rent-seeking offers a useful framework for analyzing this problem. Economic agents may allocate scarce resources toward obtaining preferential access to rents rather than engaging in productive investment and innovation (Tullock, 1967; Krueger, 1974; Bhagwati, 1982; Muradov, 2022a). In resource-dependent economies, where governments often control significant resource revenues, competition for rents can become an important determinant of economic performance. Previous studies have shown that strategic interactions associated with rent-seeking behavior may influence resource allocation and economic outcomes in important ways (Lane and Tornell, 1996; Tornell and Lane, 1999; Robinson et al., 2006).

Against this background, the present study examines how resource wealth influences the allocation of economic effort between productive activities and rent-seeking behavior. In particular, the analysis investigates whether increasing resource rents encourage economic agents to shift effort away from production and toward the acquisition of resource-generated benefits, and how such behavior may affect structural transformation. By focusing on the incentives created by resource wealth, the paper seeks to provide a complementary explanation for the persistent diversification challenges observed in many resource-rich economies.

The study contributes to the literature in several ways. First, it develops a theoretical framework in which rent-seeking emerges as an endogenous response to resource abundance rather than merely an institutional outcome. Second, it highlights the role of economic incentives in determining how effort is allocated between productive and unproductive activities. Third, it links the allocation of economic effort to broader questions of structural transformation and economic development. In doing so, the paper builds upon existing research on natural resources, rent-seeking, and

economic structure while offering a new perspective on the mechanisms through which resource wealth may influence long-run development outcomes (Muradov, 2023; Muradov and Takase, 2023).

The remainder of the paper is organized as follows. Section 2 reviews the related literature. Section 3 presents the theoretical framework. Section 4 derives the equilibrium results and discusses the main propositions. Section 5 considers the implications of the model for structural transformation, and Section 6 concludes.

Literature Review

The relationship between natural resources and economic development has generated extensive debate in both theoretical and empirical economics. Traditional economic thinking viewed natural resources as a source of comparative advantage capable of accelerating industrialization and economic growth. However, subsequent research revealed that resource abundance does not always produce favorable development outcomes. This observation led to the emergence of the resource curse literature, which argues that resource-rich countries frequently experience slower growth, weaker institutions, and limited economic diversification despite substantial resource revenues (Auty, 1993; Sachs and Warner, 1995; Gylfason, 2001; Frankel, 2010; van der Ploeg, 2011).

One of the most influential explanations for these outcomes is provided by the Dutch disease framework. Corden and Neary (1982) demonstrated that a resource boom may affect economic structure through both resource movement and spending effects. The expansion of the resource sector attracts labor and capital away from manufacturing activities, while higher income generated by resource exports increases domestic demand and appreciates the real exchange rate. As a consequence, the competitiveness of non-resource tradable sectors declines. Subsequent studies extended this framework by emphasizing investment dynamics and long-run productivity effects. van Wijnbergen (1984) argued that temporary resource booms may discourage investment in manufacturing, while Krugman (1987) highlighted the potential loss of learning-by-doing effects associated with industrial decline (Corden, 1984; Neary and van Wijnbergen, 1986; Torvik, 2001; van der Ploeg and Venables, 2013).

Although Dutch disease models successfully explain several important mechanisms linking resource abundance to structural transformation, they primarily focus on market adjustments and relative price movements. These approaches generally assume that economic agents allocate resources efficiently and pay limited attention to the incentives created by resource-generated rents. Yet resource-rich economies often exhibit substantial competition for access to resource revenues, suggesting that the allocation of economic effort itself may be influenced by resource abundance (Auty, 2001; Baland and Francois, 2000; Torvik, 2002).

The theory of rent-seeking provides a useful framework for examining this issue. Tullock (1967) first formalized the idea that individuals may expend resources to obtain transfers rather than create new wealth. Krueger (1974) further demonstrated that government-created rents can encourage unproductive competition among economic agents. In such environments, resources that could otherwise be devoted to production, innovation, and investment are instead allocated toward securing privileged access to existing wealth. From this perspective, economic inefficiency arises not only from market distortions but also from strategic behavior aimed at capturing rents (Bhagwati, 1982; Buchanan et al., 1980; Muradov, 2022a).

The interaction between natural resources and rent-seeking has attracted increasing attention within the resource economics literature. Lane and Tornell (1996) developed the concept of the voracity effect, arguing that resource windfalls may intensify competition among powerful groups and lead to excessive appropriation of economic rents. Tornell and Lane (1999) further demonstrated that weak institutional environments can amplify these effects, generating inefficient allocation outcomes. Similarly, Mehlum et al. (2006) emphasized that institutional quality determines whether resource wealth is directed toward productive activities or rent extraction. Their analysis suggests that resource abundance becomes problematic when economic agents find rent-seeking more profitable than productive entrepreneurship (Torvik, 2009; Acemoglu et al., 2004).

The broader political economy literature reaches similar conclusions. Robinson et al. (2006) argue that resource revenues may alter incentives faced by governments and interest groups, encouraging redistribution and patronage rather than productive investment. Acemoglu and Robinson (2012) likewise emphasize that the economic consequences of resource abundance depend on the institutional framework governing access to rents. These studies highlight the importance of incentives and strategic interactions in determining how resource wealth influences economic development (North et al., 2009; Collier and Hoeffler, 2005).

More recent theoretical research has examined the relationship between natural resources, market structure, and rent appropriation mechanisms. Contributions by Muradov (2021b) emphasize the importance of natural resources for economic development and introduce strategic rent-seeking behavior into resource-dependent economies. Subsequent studies further develop this line of research by examining monopoly structures, market concentration, and collusive rent-seeking arrangements under resource abundance (Muradov, 2022b, 2022c). These analyses demonstrate that strategic interactions among economic agents can influence the allocation of resource-generated rents and affect broader economic outcomes. More recent work extends the discussion by incorporating income effects, resource allocation mechanisms, and alternative preference structures within rent-seeking environments (Muradov, 2023; Muradov and Takase, 2023). Collectively, this body of research highlights the importance of incentive structures and strategic behavior in understanding the economic consequences of resource wealth.

Despite these contributions, an important gap remains in the literature. Existing studies generally focus either on the macroeconomic consequences of resource booms or on rent extraction within specific market structures. Less attention has been devoted to the allocation of economic effort between productive activities

and rent-seeking behavior when resource wealth creates new opportunities for rent acquisition. In particular, the mechanism through which resource abundance changes the relative attractiveness of production and rent-seeking remains insufficiently explored. Addressing this issue is important because structural transformation ultimately depends on how economic agents allocate their scarce resources and effort across competing activities.

The present study contributes to this literature by developing a theoretical framework in which resource wealth influences economic outcomes through its effect on the allocation of economic effort. By modeling the choice between productive activity and rent-seeking behavior, the analysis provides a complementary explanation for persistent diversification challenges observed in many resource-rich economies.

Theoretical Framework

Consider a small open economy endowed with natural resources. The economy consists of a productive sector and a resource sector. Resource wealth generates economic rents that are distributed through institutional arrangements. Economic agents possess a fixed amount of effort that can be allocated either to productive activities or to rent-seeking activities aimed at obtaining a share of resource-generated rents.

Let total effort be normalized to one:

$$E = 1$$

Each representative agent allocates effort between productive activity (p) and rent-seeking activity (r), such that:

$$p + r = 1$$

where:

$p \geq 0$ denotes productive effort,

$r \geq 0$ denotes rent-seeking effort.

Productive output is generated according to:

$$Y = A \cdot p$$

where $A > 0$ is a productivity parameter.

The economy possesses an exogenous stock of resource rents denoted by:

$$R > 0$$

Economic agents compete for these rents through rent-seeking activities. Following the contest literature, the share of resource rents captured by agent i is determined by:

$$s_i = r_i / \Sigma r_j$$

where Σr_j represents the total rent-seeking effort exerted by all agents.

The rent income obtained by agent i is:

$$B_i = s_i R$$

Substituting the contest success function into the rent-income equation gives:

$$B_i = (r_i / \Sigma r_j) R$$

Total income of agent i consists of productive income and rent income:

$$\pi_i = Ap_i + (r_i / \Sigma r_j) R$$

The representative agent chooses the allocation of effort between productive and rent-seeking activities in order to maximize total income.

Using the effort constraint:

$$p_i = 1 - r_i$$

the objective function can be rewritten as:

$$\pi_i = A(1 - r_i) + (r_i / \Sigma r_j) R$$

The optimization problem therefore becomes:

$$\text{Max } \pi_i = A(1 - r_i) + (r_i / \Sigma r_j) R$$

subject to:

$$0 \leq r_i \leq 1$$

The equilibrium level of rent-seeking effort is obtained by solving the first-order condition with respect to r_i. Assuming a symmetric equilibrium in which all agents choose identical rent-seeking effort levels, the model derives the equilibrium allocation of effort between productive and rent-seeking activities.

Optimization and Equilibrium Derivation

The representative agent chooses the level of rent-seeking effort in order to maximize total income. Using the effort constraint,

$$p_i = 1 - r_i$$

the objective function becomes:

$$\pi_i = A(1 - r_i) + (r_i / \Sigma r_j) R$$

Assume that there are n identical agents competing for resource rents. Let total rent-seeking effort be:

$$\Sigma r_j = r_i + \Sigma r_{-i}$$

where Σr_{-i} denotes the rent-seeking effort of all other agents.

Substituting this expression into the objective function yields:

$$\pi_i = A(1 - r_i) + [r_i / (r_i + \Sigma r_{-i})] R$$

The representative agent chooses r_i to maximize π_i. Taking the first derivative with respect to r_i gives:

$$\partial \pi_i / \partial r_i = -A + R[(\Sigma r_{-i}) / (r_i + \Sigma r_{-i})^2]$$

Setting the first-order condition equal to zero:

$$-A + R[(\Sigma r_{-i}) / (r_i + \Sigma r_{-i})^2] = 0$$

Rearranging:

$$A = R[(\Sigma r_{-i}) / (r_i + \Sigma r_{-i})^2]$$

To obtain a symmetric Nash equilibrium, assume that all agents choose the same level of rent-seeking effort:

$$r_i = r$$

Under symmetry:

$$\Sigma r_{-i} = (n - 1)r$$

and

$$r_i + \Sigma r_{-i} = nr$$

Substituting these expressions into the first-order condition:

$$A = R[(n - 1)r] / (nr)^2$$

Simplifying:

$$A = R(n - 1) / (n^2 r)$$

Solving for the equilibrium level of rent-seeking effort:

$$r^* = R(n - 1) / (An^2)$$

The equilibrium level of productive effort is obtained from the effort constraint:

$$p^* = 1 - r^*$$

Substituting the equilibrium value of rent-seeking effort:

$$p^* = 1 - R(n - 1) / (An^2)$$

Equilibrium productive output is therefore:

$$Y^* = Ap^*$$

Substituting p^*:

$$Y^* = A[1 - R(n - 1) / (An^2)]$$

which simplifies to:

$$Y^* = A - R(n - 1) / n^2$$

The equilibrium solutions indicate that economic agents allocate effort between productive and rent-seeking activities according to the size of resource rents, the productivity of the economy, and the number of competing agents. An increase in resource rents

raises the attractiveness of rent-seeking activities, while higher productivity increases the returns to productive effort.

Equilibrium Results and Propositions

Using the equilibrium solution

$$r^* = R(n-1)/(An^2),$$

the main properties of the model can be established.

Proposition 1. An increase in resource wealth increases equilibrium rent-seeking effort.

Proof.

Differentiating the equilibrium rent-seeking function with respect to resource rents R:

$$\partial r^*/\partial R = (n-1)/(An^2)$$

Since $A > 0$ and $n > 1$,

$$\partial r^*/\partial R > 0$$

Therefore, an increase in resource wealth raises equilibrium rent-seeking effort.

Q.E.D.

Proposition 2. An increase in resource wealth reduces productive effort.

Proof.

The equilibrium level of productive effort is

$$p^* = 1 - R(n-1)/(An^2)$$

Differentiating with respect to R:

$$\partial p^*/\partial R = -(n-1)/(An^2)$$

Since $A > 0$ and $n > 1$,

$$\partial p^*/\partial R < 0$$

Therefore, an increase in resource wealth reduces productive effort.

Q.E.D.

Proposition 3. An increase in resource wealth reduces productive output.

Proof.

The equilibrium level of output is

$$Y^* = A - R(n-1)/n^2$$

Differentiating with respect to R:

$$\partial Y^*/\partial R = -(n-1)/n^2$$

Since $n > 1$,

$$\partial Y^*/\partial R < 0$$

Therefore, higher resource wealth reduces productive output through the reallocation of effort toward rent-seeking activities.

Q.E.D.

Proposition 4. Higher productivity reduces equilibrium rent-seeking effort.

Proof.

Differentiating the equilibrium rent-seeking function with respect to A:

$$\partial r^*/\partial A = -R(n-1)/(A^2n^2)$$

Since $R > 0$, $A > 0$, and $n > 1$,

$$\partial r^*/\partial A < 0$$

Therefore, higher productivity reduces incentives for rent-seeking behavior.

Q.E.D.

Proposition 5. Higher productivity increases productive effort.

Proof.

Differentiating p^* with respect to A:

$$\partial p^*/\partial A = R(n-1)/(A^2n^2)$$

Since $R > 0$, $A > 0$, and $n > 1$,

$$\partial p^*/\partial A > 0$$

Therefore, higher productivity increases productive effort.

Q.E.D.

Proposition 6. Higher productivity increases equilibrium output.

Proof.

Differentiating Y^* with respect to A:

$$\partial Y^*/\partial A = 1$$

Since

$$\partial Y^*/\partial A > 0,$$

higher productivity increases equilibrium output.

Q.E.D.

The propositions demonstrate that resource wealth and productivity affect economic performance through their influence on the allocation of economic effort. While larger resource rents encourage rent-seeking activities and reduce productive effort, higher productivity strengthens incentives for production and improves economic outcomes. These results suggest that the economic consequences of resource abundance depend not only on the availability of resource wealth but also on the relative returns to productive activities.

Discussion and Implications for Structural Transformation

The model demonstrates that resource wealth influences economic outcomes by affecting the allocation of effort between productive and rent-seeking activities. Unlike conventional explanations that emphasize exchange rate appreciation or factor reallocation, the present framework highlights how resource rents alter the incentives faced by economic agents. As resource wealth increases, the potential returns from capturing a share of available rents become larger, encouraging agents to devote more effort to rent-seeking activities. Consequently, productive effort declines, leading to lower output and weaker structural transformation.

The results provide an alternative perspective on the challenges frequently observed in resource-rich economies. Traditional Dutch disease models emphasize the movement of labor and capital toward the booming resource sector. In contrast, the current model suggests that economic effort may also be diverted away from productive activities toward competition for resource-generated rents. Thus, the adverse effects of resource wealth need not arise solely from changes in relative prices or factor allocation. They may also emerge because economic agents increasingly focus on obtaining existing wealth rather than creating new wealth.

The equilibrium results indicate that larger resource rents increase rent-seeking incentives and reduce productive effort. This finding is consistent with the broader rent-seeking literature, which argues that economic agents respond to incentives created by available rents (Tullock, 1967; Krueger, 1974). When resource revenues become sufficiently large, competition for these rents may become more attractive than productive investment. As a result, resources that could otherwise contribute to innovation, entrepreneurship, and industrial development are redirected toward rent appropriation activities.

The model also highlights the importance of productivity. Higher productivity increases the return to productive activities and reduces the relative attractiveness of rent-seeking behavior. In economies where

productive sectors offer strong opportunities for investment and innovation, economic agents have greater incentives to allocate effort toward wealth creation rather than rent extraction. Consequently, productivity-enhancing policies may play an important role in mitigating the adverse effects associated with resource dependence.

The implications for structural transformation are particularly important. Structural transformation generally involves the reallocation of resources toward more productive sectors, technological upgrading, and economic diversification. However, the model suggests that abundant resource wealth may weaken this process by encouraging effort allocation toward rent-seeking activities. Even when resource revenues generate substantial income, long-run development may be constrained if economic agents increasingly compete for existing rents instead of investing in productive capacity. In this sense, the allocation of economic effort becomes a critical determinant of development outcomes.

These findings complement previous studies that emphasize the relationship between natural resources, rent-seeking, and economic performance (Lane and Tornell, 1996; Mehlum, Moene, and Torvik, 2006; Robinson, Torvik, and Verdier, 2006). They also extend recent theoretical research on resource rents and strategic behavior by focusing explicitly on the allocation of effort between productive and unproductive activities (Muradov, 2021b; Muradov, 2022b; Muradov, 2023). Rather than treating rent-seeking as merely an institutional failure, the model shows how such behavior can emerge endogenously from the incentives created by resource wealth.

From a policy perspective, the results suggest that resource-rich economies should focus not only on managing resource revenues but also on strengthening incentives for productive activities. Policies that improve productivity, encourage entrepreneurship, support technological development, and reduce opportunities for rent appropriation may help ensure that resource wealth contributes to long-run economic transformation. Institutional arrangements that limit the private benefits associated with rent-seeking can further enhance the attractiveness of productive investment and facilitate diversification.

Overall, the analysis suggests that the economic consequences of resource wealth depend crucially on how economic agents allocate their effort. Resource abundance does not necessarily hinder development; however, when the rewards from rent-seeking exceed the rewards from productive activity, structural transformation may be weakened. Understanding this allocation mechanism provides a useful framework for explaining why some resource-rich economies achieve diversification while others remain dependent on resource-based rents.

Conclusion

This paper develops a theoretical framework to examine how resource wealth affects the allocation of economic effort between productive activities and rent-seeking behavior. The model demonstrates that higher resource rents increase incentives for rent-seeking activities and reduce productive effort, which may weaken productive capacity and slow structural transformation.

The study contributes to the literature on natural resources and economic development by identifying the allocation of economic effort as an additional mechanism through which resource wealth influences economic outcomes. Unlike traditional explanations that focus mainly on price effects, factor movement, and institutional factors, this paper emphasizes how resource-generated rents change the relative attractiveness of productive and rent-seeking activities.

The results show that resource abundance may encourage competition for existing rents rather than productive investment, while higher productivity strengthens incentives for productive activities and reduces rent-seeking behavior. These findings highlight the importance of policies that improve productivity, encourage innovation, and increase the returns to productive investment in resource-rich economies.

Future research may extend the framework by incorporating institutional quality, heterogeneous agents, government behavior, and dynamic investment decisions. Overall, the paper demonstrates that the economic impact of resource wealth depends not only on the availability of resources but also on how economic agents allocate their effort between productive and rent-seeking activities.

References

1. Acemoglu, D., Robinson, J. A., & Verdier, T. (2004). Kleptocracy and divide-and-rule: A model of personal rule. *Journal of the European Economic Association, 2*(2–3), 162–192. https://doi.org/10.1162/154247604323067916

2. Acemoglu, D., & Robinson, J. A. (2012). *Why nations fail: The origins of power, prosperity, and poverty*. Crown Publishers.

3. Auty, R. M. (1993). *Sustaining development in mineral economies: The resource curse thesis*. Routledge.

4. Auty, R. M. (Ed.). (2001). *Resource abundance and economic development*. Oxford University Press.

5. Baland, J.-M., & Francois, P. (2000). Rent-seeking and resource booms. *Journal of Development Economics, 61*(2), 527–542. https://doi.org/10.1016/S0304-3878(00)00067-4

6. Bhagwati, J. N. (1982). Directly unproductive profit-seeking (DUP) activities. *Journal of Political Economy, 90*(5), 988–1002. https://doi.org/10.1086/261104

7. Buchanan, J. M., Tollison, R. D., & Tullock, G. (Eds.). (1980). *Toward a theory of the rent-seeking society*. Texas A&M University Press.

8. Collier, P., & Hoeffler, A. (2005). Resource rents, governance, and conflict. *Journal of Conflict Resolution, 49*(4), 625–633. https://doi.org/10.1177/0022002705277551

9. Corden, W. M. (1984). Booming sector and Dutch disease economics: Survey and consolidation. *Oxford Economic Papers, 36*(3), 359–380. https://doi.org/10.1093/oxfordjournals.oep.a041643

10. Corden, W. M., & Neary, J. P. (1982). Booming sector and de-industrialisation in a small open economy. *The Economic Journal, 92*(368), 825–848. https://doi.org/10.2307/2232670

11. Frankel, J. A. (2010). The natural resource curse: A survey. *National Bureau of Economic Research Working Paper Series*, No. 15836. https://doi.org/10.3386/w15836

12. Gylfason, T. (2001). Natural resources, education, and economic development. *European Economic Review, 45*(4–6), 847–859. https://doi.org/10.1016/S0014-2921(01)00127-1

13. Krueger, A. O. (1974). The political economy of the rent-seeking society. *American Economic Review, 64*(3), 291–303.

14. Krugman, P. (1987). The narrow moving band, the Dutch disease, and the competitive consequences of Mrs. Thatcher: Notes on trade in the presence of dynamic scale economies. *Journal of Development Economics, 27*(1–2), 41–55. https://doi.org/10.1016/0304-3878(87)90005-8

15. Lane, P. R., & Tornell, A. (1996). Power, growth, and the voracity effect. *Journal of Economic Growth, 1*, 213–241. https://doi.org/10.1007/BF00138863

16. Mehlum, H., Moene, K., & Torvik, R. (2006). Institutions and the resource curse. *The Economic Journal, 116*(508), 1–20. https://doi.org/10.1111/j.1468-0297.2006.01045.x

17. Muradov, A. (2021a). The importance of natural resources for the Azerbaijani economy. *Economics, Business and Organization Research, 3*(1), 117–131.

18. Muradov, A. (2021b, December 4–5). Natural resources and Stackelberg rent-seeking in the context of Dutch disease [Conference presentation abstract]. *19th Asia Pacific Conference: Diversity and Inclusion: Program and Abstract Book* (p. 87),

19. Ritsumeikan Asia Pacific University, Beppu, Japan. https://en.apu.ac.jp/apconf/program/2021/Program_and_abstract_book_finalversion_E.pdf

20. Muradov, A. (2022a). Energy sector's monopoly rent-seeking and supplementary time-series analysis in the case of Azerbaijan. *Anadolu University Journal of Faculty of Economics, 4*(1), 36–49. https://doi.org/10.54737/aujfe.1065684

21. Muradov, A. (2022b). Natural resources and rent seeking collusion in the context of Dutch disease. *Istanbul Journal of Economics, 72*(2), 599–623. https://doi.org/10.26650/ISTJECON2022-1063190

22. Muradov, A. (2022c, December). *Natural resources, and a rent-seeking monopoly in the context of Dutch disease* [Paper presentation]. International Conference of "Sino-Japanese Graduate Seminar on Economics and Management", Jinan University (China) and Fukuoka University (Japan).

23. Muradov, A. (2023). El efecto de los recursos naturales, el ingreso nacional y el movimiento de recursos en un monopolio de búsqueda de rentas. *El Trimestre Económico, 90*(357), 119–154. https://doi.org/10.20430/ete.v90i357.1486

24. Muradov, A., & Takase, M. (2023). *Rent seeking monopoly with VES utility function*. Center for Advanced Economic Study Working Paper Series, WP-2023-002, Fukuoka University. https://www.econ.fukuoka-u.ac.jp/researchcenter/workingpapers/WP-2023-002.pdf

25. Neary, J. P., & van Wijnbergen, S. (Eds.). (1986). *Natural resources and the macroeconomy*. MIT Press.

26. North, D. C., Wallis, J. J., & Weingast, B. R. (2009). *Violence and social orders: A conceptual framework for interpreting recorded human history*. Cambridge University Press.

27. Robinson, J. A., Torvik, R., & Verdier, T. (2006). Political foundations of the resource curse. *Journal of Development Economics, 79*(2), 447–468. https://doi.org/10.1016/j.jdeveco.2006.01.008

28. Sachs, J. D., & Warner, A. M. (1995). Natural resource abundance and economic growth. *National Bureau of Economic Research Working Paper Series*, No. 5398. https://doi.org/10.3386/w5398

29. Tornell, A., & Lane, P. R. (1999). The voracity effect. *American Economic Review, 89*(1), 22–46. https://doi.org/10.1257/aer.89.1.22

30. Torvik, R. (2001). Learning by doing and the Dutch disease. *European Economic Review, 45*(2), 285–306. https://doi.org/10.1016/S0014-2921(00)00050-8

31. Torvik, R. (2002). Natural resources, rent seeking and welfare. *Journal of Development Economics, 67*(2), 455–470. https://doi.org/10.1016/S0304-3878(01)00195-X

32. Tullock, G. (1967). The welfare costs of tariffs, monopolies, and theft. *Western Economic Journal, 5*(3), 224–232. https://doi.org/10.1111/j.1465-7295.1967.tb01923.x

33. van der Ploeg, F. (2011). Natural resources: Curse or blessing? *Journal of Economic Literature, 49*(2), 366–420. https://doi.org/10.1257/jel.49.2.366

34. van der Ploeg, F., & Venables, A. J. (2013). Natural resource wealth: The challenge of managing a windfall. *Annual Review of Economics, 4*, 315–337. https://doi.org/10.1146/annurev-economics-080511-111003

35. van Wijnbergen, S. (1984). The "Dutch disease": A disease after all? *The Economic Journal, 94*(373), 41–55. https://doi.org/10.2307/2232214

GEOLOGICAL AND MINERALOGICAL SCIENCES

LITHOLOGICAL-PETROGRAPHIC DESCRIPTION OF ROCKS OF BOKANTOV MOUNTAINS (KOKPATAS SUITE)

Boboyorov Kh.
Senior Researcher, Department of Engineering Seismology
Seismology Institute of Sciences, AS RUz, Tashkent, Uzbekistan,
https://orcid.org/0009-0000-5260-1375
Rakhmatov A.
PhD, Doctor of Philosophy in Physical and Mathematical Sciences,
Seismology Institute of Sciences, AS RUz, Tashkent, Uzbekistan,
https://orcid.org/0009-0003-3065-1824
Qodirov M.
Senior Researcher, Department of Engineering Seismology
Seismology Institute of Sciences, AS RUz, Tashkent, Uzbekistan,
https://orcid.org/0009-0005-9457-5482
Mukhammadkulov N.
Full-time doctoral researcher
Seismology Institute of Sciences, AS RUz, Tashkent, Uzbekistan,
https://orcid.org/0009-0005-9457-5482,
Fazliddinova M.
Senior Researcher, Department of Engineering Seismology
Seismology Institute of Sciences, AS RUz, Tashkent, Uzbekistan,
https://orcid.org/0009-0003-7665-3729
DOI: 10.5281/zenodo.22107132

ABSTRACT

The Bukantau Mountains, located within the northern part of the South Tien Shan orogenic belt, host numerous gold and gold-rare-metal deposits genetically associated with granitoid intrusions. This study investigates the lithological and petrographic characteristics of the host rocks to improve the understanding of their metamorphic evolution and their significance for ore localization.

Detailed petrographic investigations, supported by microscopic analyses, reveal that the regional metamorphic sequence is dominated by volcaniclastic sandstones, siltstones, slates, carbonate rocks, and associated pyroclastic deposits belonging to the greenschist facies.

Contact metamorphic rocks, including hornfels and hornfelsic slates, occur as zonally distributed aureoles surrounding granitoid intrusions and are represented by amphibolite-hornfels and muscovite-hornfels facies. Petrographic observations indicate widespread recrystallization, chloritization, sericitization, silicification, and localized sulfide mineralization dominated by pyrite and pyrrhotite.

Carbonaceous-rich horizons display enhanced sulfide enrichment, suggesting a close relationship between regional metamorphism, hydrothermal processes, and ore formation. The mineralogical composition, textural features, and spatial distribution of the lithological units demonstrate that the host rocks preserve clear evidence of both regional and contact metamorphism. These characteristics provide valuable geological and petrographic criteria for identifying concealed granitoid intrusions and assessing the exploration potential for gold and rare-metal mineralization in the Bukantau ore district.

The analysis of elemental abundances, mean concentrations, and enrichment coefficients in the host rocks of the Bukantau Mountains demonstrates that gold mineralization is strongly controlled by the lithological and petrographic composition of the enclosing rocks. The highest gold concentrations are predominantly associated with fractured and silicified sandstone-siltstone units. Hydrothermal fluids migrating through fracture networks facilitated the precipitation of quartz-vein mineralization, leading to the formation of vein-type ore bodies.

Keywords: Kyzylkum, regional, chlorite, sericite, facies, silicon, limestone, calcite, exocontact, hornfels, biotite, contact, Caledonian, phthanites, stratum, microgranular.

INTRODUCTION

The present study focuses on the lithological and petrographic characterization of the host rocks exposed in the Bukantau Mountains and their genetic relationship with gold and rare-metal mineralization. Particular attention is given to the mineralogical composition, textural and structural attributes, metamorphic evolution, and ore-related alteration processes that control the distribution of mineralization. The results establish new petrographic indicators that may serve as effective exploration criteria for concealed ore-bearing systems and contribute to a more comprehensive understanding of the geological evolution of the Bukantau ore district.

Detailed investigation of the lithological, mineralogical, and petrographic characteristics of the host rocks constitutes a fundamental component of modern mineral exploration. Such studies are especially important in newly prospective areas, where recognition of diagnostic mineral assemblages and metamorphic

transformations provides critical constraints on ore-forming processes. Consequently, these data improve geological targeting, facilitate the delineation of prospective mineralized zones, and provide a scientific basis for the efficient exploration and sustainable development of gold and rare-metal deposits.

Literature review. The initial geological studies of the area were primarily devoted to investigating its geological structure, stratigraphy, tectonic setting, and magmatic evolution. Fundamental contributions to the understanding of these geological aspects were made by I.P. Gerasimov, P.K. Chikhachev, A.F. Sosedko, N.P. Petrov, K.K. Pyatkov, A.K. Bukharin, Kh.R. Rakhmatullaev, V.I. Zonov, A.A. Shapkin, A.A. Rubanov, S.Y. Lapidus, L.M. Komarova, Yu.T. Abramov, P.R. Khakimdjanov, N.L. Kholopov, S.A. Kushnar, Ya.A. Likhachev, M.A. Akhmedjanov, E.R. Bozorboev, A.Y. Kotunov, V.S. Karsakov, T.Sh. Shayakubov, T.N. Dalimov, Ya.B. Aysanov, Z.M. Abduazimova, Yu.G. Abdullaev, and numerous other geologists, whose investigations significantly advanced the geological knowledge of the region.

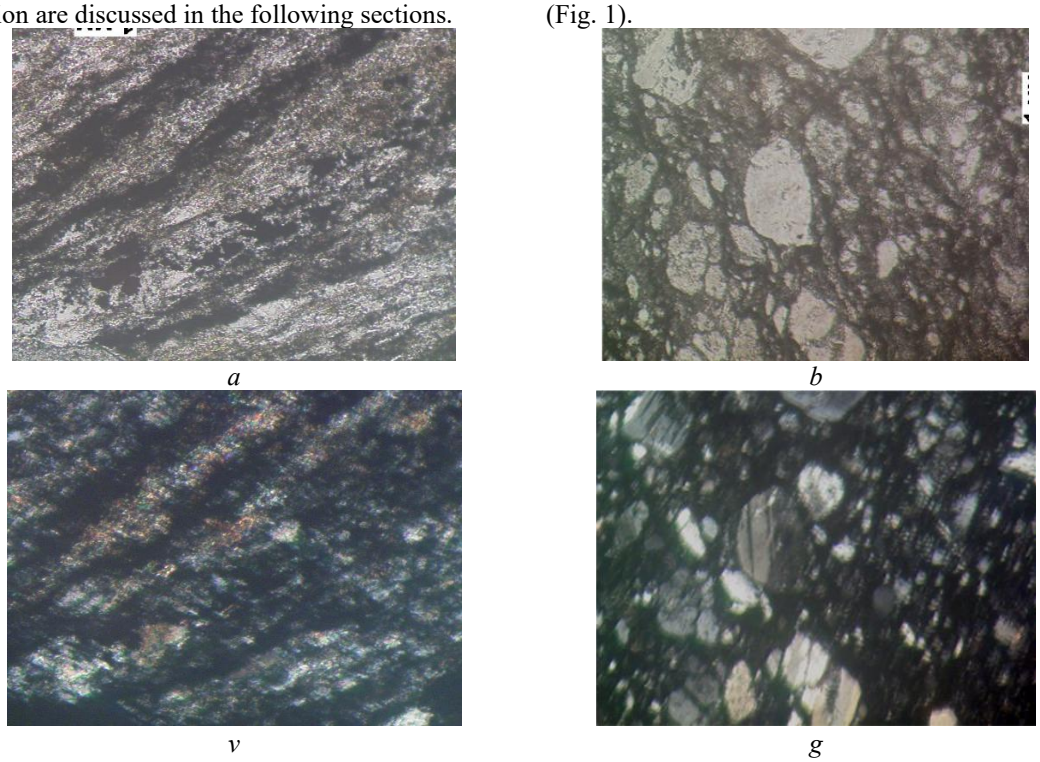
The Bukantau Mountains form part of the linear cover-fold belt of the Southern Tien Shan and occupy its northern structural domain [1]. From a tectonic perspective, the study area consists of a Lower Caledonian-Hercynian folded basement complex unconformably overlain by the younger Alpine structural sequence. The sedimentary succession comprises deposits accumulated within the continental slope, continental margin, and shelf environments along the passive margin of the Kyzylkum-Nurata microcontinent [2]. To characterize these geological units, comprehensive petrographic investigations were carried out using optical microscopy, and the lithological and petrographic features identified during the microscopic examination are discussed in the following sections.

The Bukantau ore district comprises several economically significant gold-rare-metal, gold, and gold-silver deposits. Gold-rare-metal mineralization is predominantly localized within the contact zones of granitoid intrusions, whereas the gold and gold-ilver deposits are mainly controlled by tectonic fault systems spatially associated with dyke intrusions [3,4].

Research Methodology. Based on a detailed investigation of the genesis and petrographic characteristics of the rocks within the study area, the principal lithological units belonging to the regional metamorphic and contact metamorphic complexes were distinguished. This classification forms the basis for the interpretation and theoretical synthesis of laboratory results, allowing a comprehensive assessment of the rocks structural and textural characteristics, mineralogical composition, and petrogenetic evolution.

The regional metamorphic complex is represented by rocks belonging to the chlorite-sericite facies of the greenschist metamorphic assemblage. These lithologies form the petrographic basement of the Kokpatas Formation and consist predominantly of siliceous rocks (phthanites), carbonate rocks, and compositionally diverse sandstone, siltstone, and slate.

Analysis and results. The slates are composed predominantly of an albite-chlorite-silica-sericite mineral assemblage, although their mineralogical composition varies depending on the relative proportions of the constituent minerals. They display a distinct compositional banding defined by alternating layers with thicknesses ranging from 0.1 to 1.0 mm. Individual laminae, typically less than 0.01-0.03 mm thick, are dominated by finely laminated sericite, whereas others are characterized by the predominance of chlorite. A third textural variety consists of heterogeneous micrograined aggregates composed of chlorite, sericite, and plagioclase (Fig. 1).



a *b* *v* *g*

Figure 1. Carbonaceous-chlorite-sericite schists with accumulations of (dark) carbonaceous matter along individual layers (a, v); tuffaceous sandstones composed of plagioclase fragments and a carbonaceous-chlorite groundmass (b, g); a, b – parallel nicols (II), v, g – crossed nicols (+).

The dark laminae are intensely impregnated with carbonaceous material, which accounts for approximately 2-3% of the rock volume. They contain well-preserved fragments of carbonaceous–sericite slate exhibiting pronounced schistosity and compositional layering. Petrographically, the rocks are characterized by a fine-grained lepidoblastic texture, which locally grades into an inequigranular granolepidoblastic texture within the siltstone-derived laminae. The rocks exhibit a well-developed schistose fabric accompanied by distinct banded and ribbon-like textures.

The slates are commonly characterized by disseminated sulfide mineralization, predominantly represented by pyrite, which is widely distributed throughout the rock matrix. Sulfide mineralization is preferentially concentrated within carbonaceous-rich laminae, where it exhibits greater abundance. Petrographically, pyrite occurs as euhedral to subhedral crystals with grain sizes of approximately 0.5-0.7 mm, as well as lens-shaped aggregates elongated parallel to the schistosity. In addition, isolated disseminated pyrite grains ranging from 0.01 to 1.5 mm in size are locally present.

The observed ore mineralization is interpreted to be genetically related to the recrystallization of carbonaceous strata during regional metamorphism. The sandstones and siltstones are volcaniclastic rocks derived from andesitic-basaltic pyroclastic material. Petrographic examination reveals abundant angular to subangular polygonal grains of intermediate-to basic-composition plagioclase dispersed throughout the sandstone matrix. Lithoclasts composed of andesitic-basaltic volcanic glass are commonly enclosed within, and locally cemented by, a silty-pelitic matrix. Metamorphic alteration preferentially affected the cementing groundmass, resulting in the development of fine-grained sericite and chlorite. Leucoxene occurs predominantly as irregular micro-scale aggregates within the silty-pelitic cementing matrix of the tuffaceous sandstones (Fig. 1).

Petrographically, the rocks are characterized by a medium-grained blastopsammitic texture and a fragmented framework, whereas the cement locally exhibits a lepidoblastic fabric. The rocks are generally massive, with only locally developed weak compositional layering. They are locally affected by silicification, occurring as quartz-rich lenses and veinlets, and are cross-cut by quartz-calcite veins measuring approximately 0.5-0.7 cm in thickness. The basal cement is composed predominantly of volcanic glass, with subordinate carbonaceous material.

The metasiltstones are coarse-grained volcaniclastic lithocrystalloclastic rocks composed predominantly of andesitic porphyrite lithic fragments (~45%), with subordinate albite (2-3%), pyrite (0.5-1.0%), and carbonaceous material (0.1-0.3%). Petrographic observations indicate that the plagioclase fragments are polygonal and generally angular. Owing to pervasive chloritization, the plagioclase has been extensively replaced and merged with the pelitomorphic cement, forming a homogeneous cementing matrix (Fig.2). Scattered euhedral to subhedral monocrystalline pyrite grains, reaching up to 0.2 mm in diameter, occur throughout the rock. Quartz veinlets, typically 1.0-1.5 mm thick, are locally developed.

Petrographically, the tuffaceous siltstones exhibit a blasto-siltstone texture and a weakly developed schistose fabric. Their mineralogical composition, combined with their characteristic textural and structural features, provides clear evidence of a volcaniclastic origin and supports their interpretation as pyroclastic (tuffaceous) deposits.

Carbonate rocks within the study area are represented by limestone and dolomite interbeds with thicknesses ranging from 0.5 to 8.0 m. The limestones are fine-grained and are composed predominantly of polygonal, twinned calcite crystals measuring 0.3-0.7 mm in diameter. They also contain minor disseminations of vitrinite-anthraxolite and rare pyrite crystals, generally less than 0.1 mm in size. Petrographic observations reveal irregular patchy recrystallization, producing domains composed of coarser polygonal calcite grains measuring 0.7-1.0 mm. The rocks are characterized by a fine-grained granoblastic texture and a brecciated, mottled fabric (Fig. 2).

The dolomite is medium-grained and contains scattered calcite patches. Petrographic analysis indicates that it consists of approximately 60-65% dolomite, 30-35% calcite, and 0.2-0.3% opaque pelitomorphic material, tentatively interpreted as limonite. Dolomite crystals are highly irregular in morphology, tightly interlocked, and typically range from 1.5 to 2.0 mm in diameter. Metamorphic recrystallization has locally resulted in the partial replacement of dolomite by coarse-grained calcite. The calcite crystals are characterized by well-developed polysynthetic twinning.

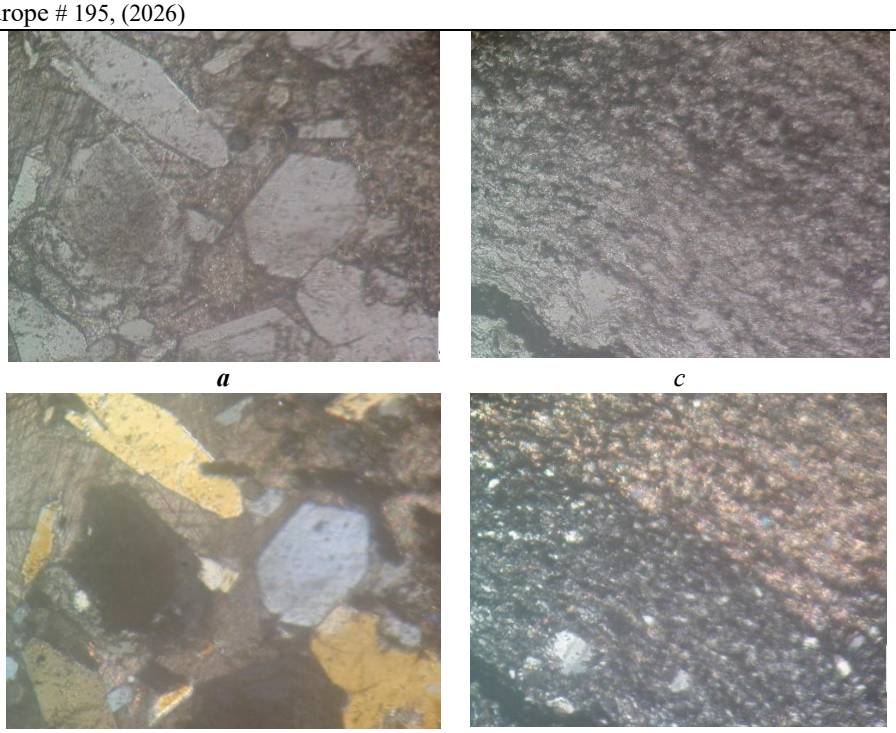


Figure 2. Calcite accumulations in recrystallization zones within fine-grained dolomites (a-v); albite-carbon-chlorite and albite-sericite-chlorite tuffaceous metasiltstones (b, g); a, b – parallel nicols (II), v, g – crossed nicols (+).

The contact metamorphic complex of the study area is developed within the exocontact aureoles surrounding the Sautbay and Saritau granitoid intrusions. Contact metamorphism extends to distances of approximately 500-600 m from the intrusive bodies and is represented by two distinct metamorphic facies: (i) an inner amphibolite-hornfels facies and (ii) an outer muscovite-hornfels facies [6-9].

Within the amphibolite-hornfels facies of contact metamorphism, high-temperature metamorphic assemblages are characterized by the paragenetic association of forsterite, clinopyroxene, andalusite, corundum, sillimanite, and spinel. In contrast, lower-temperature assemblages consist predominantly of nodular and spotted hornfelsic slates composed of albite, muscovite, and actinolite. These mineral assemblages indicate that metamorphic recrystallization corresponding to the amphibolite-hornfels facies occurred at temperatures of approximately 510-515°C.

The contact hornfels consist of varying proportions of quartz, biotite, plagioclase, and pyroxene, with their mineralogical composition being largely controlled by the composition of the protolith. Petrographically, the rocks are dark gray to greenish-gray in color and exhibit a microlepidogranoblastic texture. They are characterized by a predominantly massive fabric with locally developed mottled textures.

Petrographic analysis indicates that the hornfels consist of approximately 30-35% quartz, 30-40% plagioclase, up to 15-20% biotite, and 2-10% ore and accessory minerals, predominantly pyrite and pyrrhotite. In some lithological varieties, the groundmass is composed of nearly equal proportions of fine-grained quartz and plagioclase. Lamellar biotite forms clustered aggregates that appear as irregular patches within the fine-grained plagioclase-rich matrix, producing the characteristic mottled texture of the hornfels.

Hornfels exposed in the immediate vicinity of the granitoid intrusive contacts contain sapphirine and cordierite, which are characteristic of high-grade contact metamorphic zones. Plagioclase occurs as fine-grained tabular crystals exhibiting compositional zoning and forming aggregates ranging from 0.5 to 1.0 mm in size. Prismatic amphibole crystals contain abundant inclusions of apatite prisms (approximately 0.01 × 0.10 mm). Fine-grained titanite (sphene) is preferentially developed along fractures and grain boundaries of plagioclase crystals. Spinel is locally present as strongly corroded porphyroblasts embedded within the hornfels groundmass.

Conclusion/Recommendations. In summary, the lithological and petrographic characteristics of the host rocks in the Bukantau Mountains indicate that the regional metamorphic succession consists predominantly of compositionally diverse volcaniclastic sandstones, siltstones, and slates belonging to the greenschist facies. These rocks are accompanied by contact metamorphic assemblages, including hornfels and hornfelsic slates, which occur as zonally developed aureoles surrounding granitoid intrusions. The compositional diversity, petrographic characteristics, and spatial distribution of these lithological units constitute valuable exploration criteria for delineating concealed intrusive bodies genetically related to gold and rare-metal mineralization in the Bukantau ore district.

References

1. Biske, Yu.S., 1996. Paleozoic Structure and History of the Southern Tien Shan. Saint Petersburg University Press, Saint Petersburg, 192 p.

2. Bukharin, A.K., Brezhnev, V.D., Maslennikova, I.A., et al., 1989. Tectonics of the Western Tien Shan. Nauka Publishing House, Moscow, 152 p.

3. Ore Deposits of Uzbekistan, 2001. GIDROINGEO, Tashkent, 661 p.

4. Karabaev, M. S. (2012). Study of organic matter in the sedimentary–metamorphogenic rocks of Eastern Bukantau (Central Kyzylkum). Bulletin of the National University of Uzbekistan, (2), 46–48.

5. Yusupov, G. R., Maripova, S. T., & Yuldashev, O. A. (2003). Criteria for the prospecting and evaluation of gold mineralization of various industrial types in Southern Bukantau (Research Report No. 493, 1999–2003; 2 vols.). Institute of Mineral Resources, Tashkent, Uzbekistan.

6. Ismoilov V.A., Khusomiddinov A.S., Mukhammadkulov N.M., Yadigarov E.M., Khayriddinov B.B., Rakhmatov A.R., Jumaev D.D., Bozorov J.SH., Isroilov X.B., Mansurov A.F., Assessment of Landslide Dynamics Based on Microseismic Parameters. Seismic Instruments. Published: 20 December 2025 Volume 61, pages 346–353, (2025) https://doi.org/10.3103/S0747923925700446

7. V.A. Ismailov, A.R. Rakhmatov, A.S. Xusomiddinov, E.M. Yadigarov, J.Sh. Bozorov Assessment of the soil seismic condition through microseismic measurements (in the example of the city of Bukhara). NEWS OF THE NATIONAL ACADEMY OF SCIENCES OF THE REPUBLIC OF KAZAKHSTAN, SERIES OF GEOLOGY AND TECHNICAL SCIENCES. Volume 5. Number 473 (2025), 131–145 https://doi.org/10.32014/2025.2518-170X.554131 (скопус).

8. Jumaev D. Rakhmatov A.R. Probabilistic seismic hazard assessment (psha) for construction site in Tashkent, Uzbekistan. Sciences of Europe (Praha, Czech Republic). 2024 y. №148, pp 3-11 https://zenodo.org/records/13739249

THE S2 PARAMETER OF SHALE FORMATIONS: UNREALIZED HYDROCARBON POTENTIAL UNDER THE THERMODYNAMIC REGULATION OF FORMATION PRESSURE

Bochkarev V.
PhD
Independent Upstream consultant, Mexico.
ORCID: 0009-0004-3642-6561
Bochkarev A.
D.Sc.
Professor, Gubkin Russian State University of Oil and Gas, RF
Bochkarev Ar.
Graduate student, Gubkin Russian State University of Oil and Gas, RF
ORCID: 0009-0000-8526-3755
DOI: 10.5281/zenodo.22107163

ABSTRACT

A cyclic development scheme for shale formations with a (36 months + 3 weeks) ·N period is proposed, providing a 2–5-fold increase in cumulative production per well at capital expenditures limited to the cost of re-fracturing, with an internal rate of return above 50%. The physical basis of the scheme is a revision of classical petroleum geochemistry as applied to shale formations. First, the Rock-Eval S2 parameter is proposed to be interpreted not as a "residual" generation potential, inert on development timescales, but as the not-yet-realized hydrocarbon potential of the formation — an operational category of replenishable reserves. Second, formation pressure is treated as a thermodynamic regulator of generation (a baric brake): in a closed, self-regulating system with nanodarcy permeability, pressure buildup during generation is capped by abnormally high formation pressure (AHFP) and arrests the reaction; pressure relief by production resumes it. Third, it is shown that transport of newly generated hydrocarbons within the nanopermeable matrix is not diffusive: kerogen realization is accompanied by a 10–20% volume increase, which drives generation-induced microfracturing — the hydrocarbons create their own migration pathways; a single act of such fracturing — hydraulic fracturing via a cavitation bubble collapsing with a hydraulic shock — is documented on core from the Maykop clay shales of the Ciscaucasus. Post-refrac rate recovery is decomposed into two contributions with different characteristic timescales: a fast one — redistribution and desorption of pre-existing fluids from the SRV volume, and a slow, cumulative one (2–3 years, over N cycles) — realization of S2 under relieved pressure, manifesting as anomalous well longevity. A field experiment for direct ΔS2 measurement is proposed.

Keywords: shale formations; Rock-Eval; S2; Tmax; AHFP; baric brake; generation-induced microfracturing; hydraulic fracturing; cavitation mechanism; replenishable reserves; cyclic development; re-fracturing; Type II kerogen; Eagle Ford; Pimienta.

1. Tools calibrated before the shale revolution

The methodological apparatus of petroleum geochemistry — Rock-Eval pyrolysis, maturity scales based on Tmax and vitrinite reflectance Ro, the concepts of "realized" and "residual" generation potential — was established in the 1970s–1980s [12–14] using data from conventional petroleum systems. In these systems the source rock and the reservoir are spatially separated: generated hydrocarbons migrate and accumulate in traps, and the object of development is the reservoir, not the source. Under this framework, the S2 parameter (hydrocarbons released during programmed kerogen cracking, mg HC/g rock) is naturally treated as a residual potential — a characteristic of what the source rock could still generate over geological time, but which is irrelevant to the current development project.

The shale revolution changed the object without changing the toolkit. In formations such as Eagle Ford, Pimienta, Vaca Muerta, and the Domanik, the source rock is the reservoir: development takes place directly within the generation kitchen, at formation temperatures of 100–130 °C corresponding to the main oil generation window, and under AHFP that is a direct consequence of ongoing generation in a closed system. Nevertheless, pyrolysis is still interpreted using calibrations and thresholds established for the conventional case [14], and S2 continues to be treated as a quantity that plays no part in the reserves balance.

This paper argues the opposite position. For a shale formation, S2 is an unrealized hydrocarbon potential available for targeted engagement through development; formation pressure is a controllable thermodynamic regulator determining whether this potential is realized or not; and the transport of newly generated hydrocarbons within the nanopermeable matrix is provided not by diffusion but by the act of generation itself, through volumetric expansion and microfracturing (Figure 1). These three propositions give rise to a specific technological scheme — cyclic development with a (36 months + 3 weeks)·N period — and to a specific falsifiable experiment allowing the concept to be tested quantitatively. These propositions build on the concept of replenishable reserves formulated earlier by the authors [1, 2].

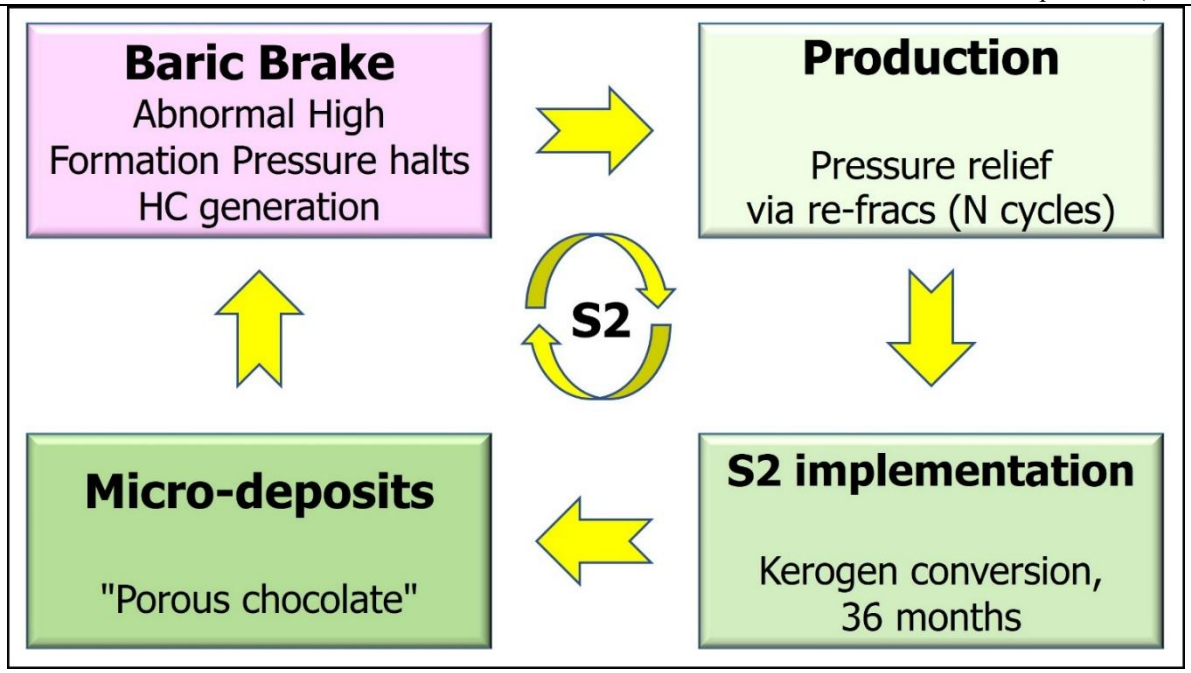


Figure 1. Principal scheme of kerogen (S2) potential realization at the onset of formation development (AHFP relief)

2. Pressure as a thermodynamic brake on generation

2.1. Closed self-regulating system

A shale formation with matrix permeability in the range of units to tens of nanodarcy is a hydrodynamically closed system: on geological timescales, fluid leakage through the caprock and base is negligible compared with the rate of generation, while on development timescales mass exchange is only possible through an artificially created hydraulic fracture network. Cracking of kerogen into liquid and gaseous hydrocarbons is accompanied by an increase in the specific volume of the products relative to the original solid organic matter. In a closed volume this means a rise in pore pressure — a generally accepted mechanism for AHFP formation in source rocks [15].

The upper limit to which pore pressure can rise within this quasi-closed cycle is set by the hydraulic fracturing pressure of the enclosing matrix (Pfr):

$$Pfr = 0.01 \cdot \upsilon/(1 - \upsilon) \cdot H \cdot \sigma_b,$$

where υ is the Poisson's ratio, H is the burial depth, and σb is the bulk density of the overlying strata [5]. The baric brake ceases to operate once formation pressure approaches this critical value (Figure 2).

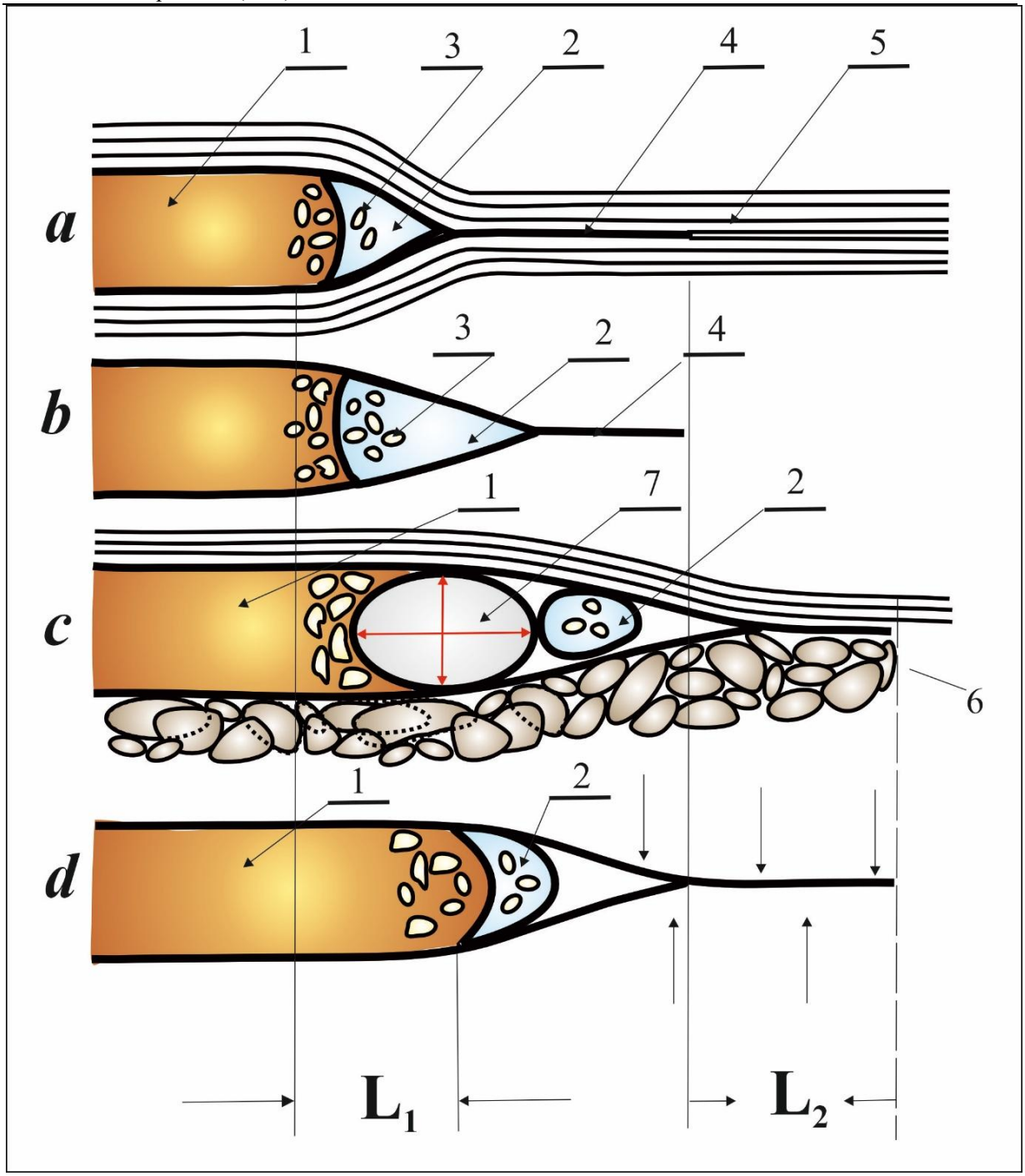


Figure 2. Mechanism of crack formation (growth): 1 — oil; 2 — water; 3 — vapor-gas nuclei and oil-water interfaces; 4 — crack walls; 5 — clay laminae; 6 — siltstone laminae; 7 — cavitation bubble (cavity): L1 and L2 are the increments of the oil micro-deposit and the crack, respectively [5]

2.2. The Le Chatelier principle and pressure-retardation empirics

For a reaction proceeding with a volume increase, rising pressure thermodynamically shifts the conditions against its progress (the Le Chatelier principle). A system generating hydrocarbons in a closed volume raises its own pressure until it reaches a state in which further generation is suppressed: AHFP acts as an automatic limiter. The formation settles into a self-regulated quasi-stationary state of a "charged" but arrested reactor. This state is proposed to be called the baric brake (pressure-controlled generation shutdown).

An objection can be anticipated: kerogen cracking is a kinetically controlled, practically irreversible process, and direct application of equilibrium thermodynamics to it calls for caution. However, suppression of organic matter maturation by excess pressure is not a speculative construct but an experimentally and empirically documented effect (pressure retardation): it has been reproduced in aqueous pyrolysis [16], described by kinetic models with a pressure correction [17], and confirmed by observations of anomalously low maturity in zones of strong AHFP in sedimentary basins [18]. Thus, regardless of the rigor of the thermodynamic analogy, the empirical fact remains that high

pore pressure retards the realization of generation potential, while its reduction removes that retardation.

2.3. Consequence for development

If pressure is the brake, fluid production is the pedal. Reducing formation pressure in the drained zone below AHFP (field data show a significant response at a depression of about 5 MPa [19]) shifts the arrested reactor into an active state: generation from S2 resumes within the volume covered by the depression. This thesis is central to the discussion that follows, and it raises two questions addressed later: how the newly generated hydrocarbons overcome the nanodarcy matrix, and on what timescale this contribution manifests.

2.4. An extreme analogue: thermal gas treatment

Controllability of generation on engineering, rather than geological, timescales has direct field confirmation in the experience of thermal gas treatment (TGT) applied to the Bazhenov formation [1, 9]. At a declared 120–130 °C, the actual temperature at the in-situ combustion front locally could reach ~180 °C; core from the affected zone was altered to a "brick-like" state, and the share of dissolved gas in production doubled — interpreted as a local shift of organic matter from the main oil generation window into the gas-condensate window. TGT is an extreme case (forcing the temperature term of the Arrhenius equation), but it demonstrates the essential point: a formation's position on the catagenesis scale is not a constant on development timescales - it can be managed. The cyclic scheme proposed below manages the baric, not the thermal, factor - more cheaply and without reservoir degradation.

2.5. Two axes for controlling generation: thermal and baric

According to the concept of two-stage generation developed earlier by one of the authors [2-4], the source rock generates predominantly liquid hydrocarbons at the first stage, and upon subsequent tectonic subsidence and transition into a higher-temperature region enters a second stage, generating gas. Two-stage generation describes the system's evolution over time — its upward movement along the catagenesis scale driven by the thermal factor.

The present work considers the sequence at a fixed point on this scale — in the maturity state and thermobaric regime in which it exists at the time of development. The question is posed differently: not where the system is moving, but whether the potential already accumulated at the level reached is being realized. This distinguishes two independent axes for controlling the generation process (Table 1).

Table 1.

Two axes for controlling hydrocarbon generation

Attribute	Thermal axis (two-stage generation)	Baric axis (present work)
What it determines	Type of fluid generated — oil, condensate, gas	The fact and rate of potential realization at the maturity level reached
Controlled by	Tectonic subsidence, heat flow	Formation pressure, and hence production and AHFP reduction
Characteristic time	Geological	Engineering – at the point development begins
Field analogue of forcing	Thermal gas treatment (TGT)	Cyclic development (re-fracturing)

Considering the thermal factor, the phenomenon of hypercatagenesis observed in the Pimienta formation in Mexico deserves particular mention. Drilling data on the Pimienta formation within the western part of the Tampico basin have established an anomalously high temperature gradient and anomalously high formation pressures in the Tithonian deposits of the Pimienta formation, resulting from the geodynamic and thermal influence of the Sierra Madre Mountain system [11]. A high-standing, viscoplastic magmatic body and its convective migration eastward caused extreme heating and accelerated catagenesis (hypercatagenesis) of organic matter (OM) in the Pimienta formation within the contact zone, starting from a depth of just 500 m at stage MK_5 (the main gas generation window), bypassing the stages of protocatagenesis (PC_{1-3}), early and middle mesocatagenesis (MC_1-4), and the main oil generation window (MC_1–MC_2). Roughly along the Gulf of Mexico coastline lies the boundary between the main gas window and the main oil window, beyond which the Pimienta formation begins to subside seaward and transitions to the conventional regional OM catagenesis scheme with a normal temperature gradient (Figure 3) [11].

The two concepts do not compete but complement each other along different axes of a single process. Two-stage generation answers the question of which hydrocarbon the sequence is capable of yielding; baric regulation answers the question of whether — and when — it will yield it. The practical consequence: the baric route is preferable as the standard development mode (gentle, cheap, non-destructive to the rock), whereas the thermal route remains a tool for extreme intervention in shale formations where hydraulic fracturing cannot be applied.

3. Transport without diffusion: generation-induced microfracturing

3.1. The feasibility of engaging S2 in development

The standard objection to the feasibility of engaging S2 in development runs as follows: at a matrix permeability on the order of nanodarcy, any fluid transport from the point of generation to a hydraulic fracture requires geological time, and consequently the newly generated hydrocarbons cannot physically reach the wellbore.

The objection is valid — and this is precisely its main weakness. If diffusive transport within the shale matrix is practically excluded, it is excluded for

any fluid, not only for newly generated ones. The objection therefore refutes not the proposed concept but the classical model explaining post-refrac rate recovery as "inflow from the matrix": inflow through that same nanodarcy matrix is equally impossible.

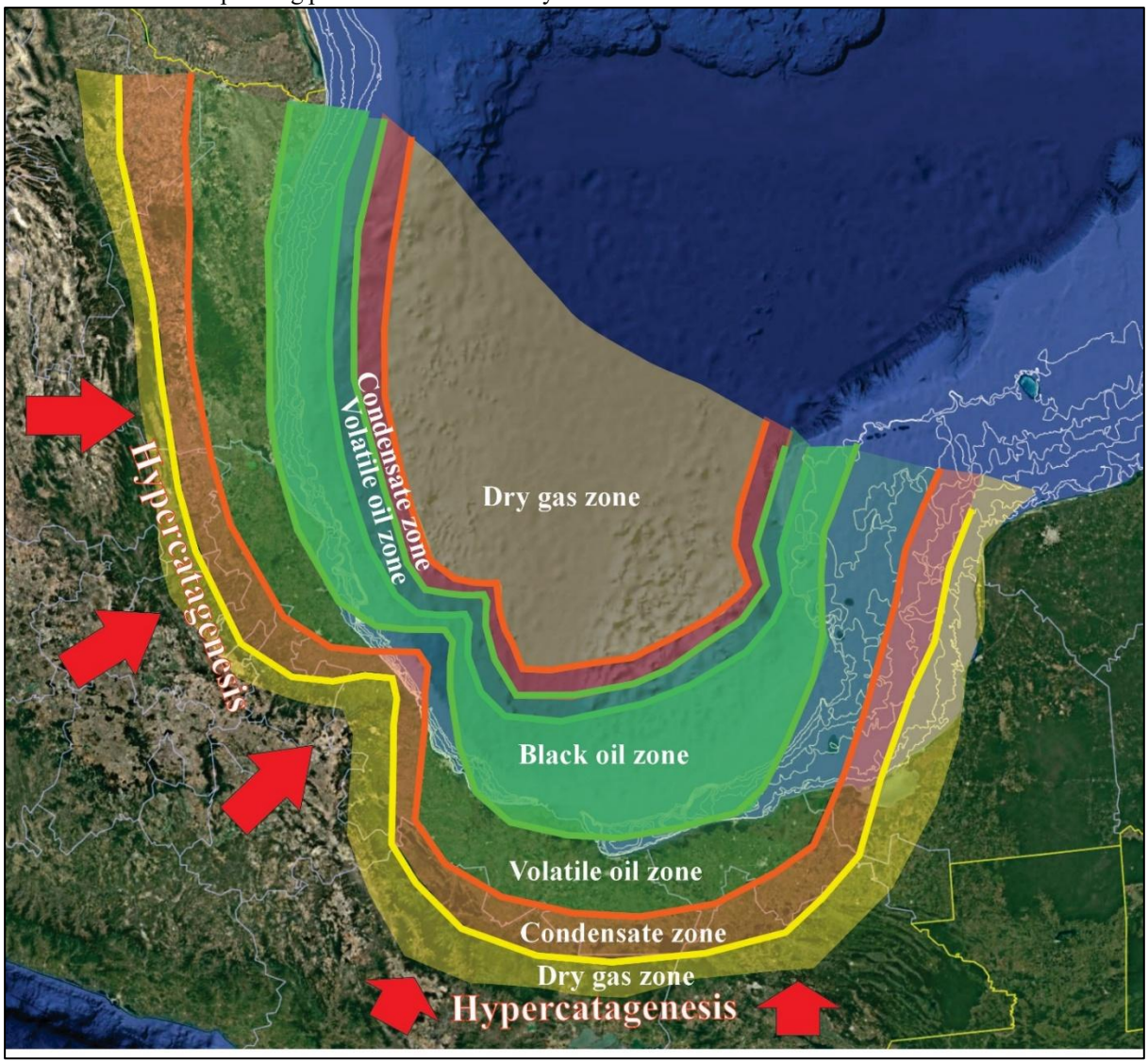


Figure 3. Catagenetic zonation of the Pimienta formation onshore and on the shelf of southeastern Mexico, showing hypercatagenesis zones (compiled by the authors from a range of open-source data).

3.2. Volumetric-baric transport mechanism

The resolution of the contradiction lies in the fact that transport of newly generated hydrocarbons is not diffusive. It is volumetric-baric: kerogen realization is accompanied by a volume increase, and this expansion itself ruptures the enclosing matrix, forming a network of microfractures.

This fracturing mechanism is documented in detail on core from the Maykop clay shales of the Central Ciscaucasus [3, 5]: crack growth proceeds through hydraulic fracturing — a cavitation bubble, forming as a water microjet detaches at the oil-rock interface, collapses with a hydraulic shock (initial pressure at the collapse center of 57–677 MPa), producing a crack increment of $5–7\times10^{-3}$ m per cycle at a speed reaching 0.3–0.7 of the Rayleigh wave velocity. The pressure limit reached before the act of hydraulic fracturing is set by the value of Pfr (Figure 2). This model, obtained from core and field material as early as 1990, predates the description of generation-induced microfracturing on Barnett and Bakken material by FIB-SEM methods by two to three decades, and provides a physical mechanism for the same phenomenon at the level of an individual fracturing event, not just its outcome.

The phenomenon is broadly consistent with independent lines of evidence obtained on other material: the accompanying volume increase during kerogen conversion to oil and gas is estimated at around 10–20% under standard conditions [20]; solid-phase expansion as a cause of microfracturing was noted as early as in the work of Pelet and Tissot [21] and Vernik [22]; and discrete-element numerical modeling together with X-ray microtomography of Green River shale show the formation of a three-dimensional percolating microfracture network connecting individual kerogen sites into a single system, nucleating at the edges of kerogen inclusions and coalescing where pressure fields from neighboring sites overlap [20, 23, 24].

The resulting structure can be described by the metaphor of "porous chocolate": a dense matrix permeated by a network of fluid-filled micro-cavities and microfractures, each representing a micro-deposit. The

newly generated hydrocarbons do not seep through the rock — they **cut their own path** (Figure 4). Transport here is mechanical rather than concentration-driven, and therefore is not subject to the constraints of diffusive kinetics.

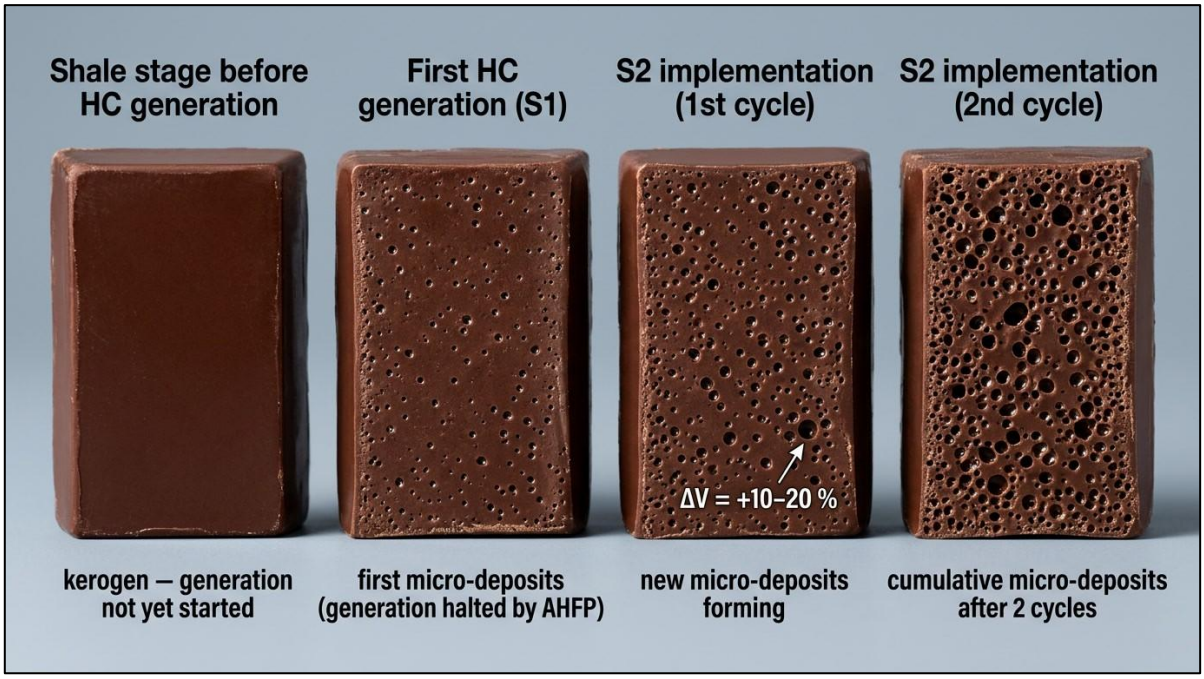

Shale stage before HC generation
First HC generation (S1)
S2 implementation (1st cycle)
S2 implementation (2nd cycle)
ΔV = +10–20 %
kerogen — generation not yet started
first micro-deposits (generation halted by AHFP)
new micro-deposits forming
cumulative micro-deposits after 2 cycles

Figure 4. Scheme of the staged realization of kerogen in the course of shale formation development — "porous chocolate".

3.3. Nano-immobility and the function of hydraulic fracturing

The resulting micro-deposits do not form a system connected to the wellbore: the percolating network is closed within the rock volume and has no outlet. This state is proposed to be called nano-immobility (sub-microdarcy transport immobility). In this framework, the function of multistage hydraulic fracturing is reinterpreted: fracturing does not "increase the reservoir permeability" in the classical sense, but rather links dispersed micro-deposits into a single drainage network connected to the wellbore. This is precisely why the effectiveness of fracturing in shale differs fundamentally from its effectiveness in conventional reservoirs, where the fracture merely accelerates inflow of an already mobile fluid.

3.4. Novelty: reproducibility of the process

The phenomenon of generation-induced microfracturing is not a discovery of this work — it has been described and observed before. However, in the existing literature it is treated as a single paleo-event: microfractures form during the main oil generation epoch and serve as primary migration pathways, after which the process is considered complete.

The novelty of the proposed approach lies in asserting the reproducibility of this process. If generation has been arrested by the baric brake rather than by kerogen exhaustion, then pressure relief through production resumes both generation and its accompanying microfracturing. Generation-induced microfracturing thereby becomes not a paleo-event but a controllable, repeatable technological process, re-triggered by every development cycle.

4. Two mechanisms of rate recovery: separating timescales

A field fact requiring explanation: after re-fracturing of wells that have completed a cycle, the rate recovers to 70–100% of the initial value, and individual Eagle Ford vertical wells have operated for more than 30 years — even though a single fracture network degrades within 1.5–2 years. It is tempting to attribute the entire recovery to renewed generation from S2. This would be a mistake from a kinetic standpoint: the cracking-rate constants of Type II kerogen at 100–130 °C (activation energies on the order of 50–60 kcal/mol [25]) yield S2 conversion rates that are clearly insufficient to replenish recoverable volumes within the first few months. A correct formulation requires separating two contributions with different characteristic timescales.

4.1. The fast contribution: redistribution and desorption of existing/undrained fluids

During the working period of a cycle, the depression cone covers the SRV volume, but only the part of it that remains hydrodynamically connected to a fracture network that still retains conductivity is actually drained. As the proppant degrades (1.5–2 years), the far-field fractures collapse, the drainage radius contracts toward the near-wellbore zone, and undrained mobile fluid remains in the far SRV zone — free hydrocarbons in the pores and microfractures of the previously formed generation network, along with hydrocarbons sorbed on kerogen and clay minerals (in pyrolysis terms, an undrained fraction of S1).

Re-fracturing reconnects this zone. Rate recovery is provided by redistribution and desorption of already existing fluids through an already existing microfracture network — filtration and mechanical processes rather than chemical ones, and therefore fast. No generation is required to explain this contribution, and the kinetic objection does not apply to it. Thus, if re-

fracturing is performed on a shale well that has been on production for 11–14 months, 60–70% of the resulting inflow will consist of previously undrained fluids.

4.2. The slow, cumulative contribution (2–3 years over N cycles): realization of S2

At the same time, in the volume where a significant part of S1 has been removed by production and pressure has dropped below AHFP, the baric brake is lifted (Section 2): conversion of S2 into mobile hydrocarbons resumes, accompanied by a new act of microfracturing (Section 3). The rate of this conversion at formation temperature is indeed low — which is precisely why its contribution is stated not on a scale of months but on the scale of inter-cycle intervals and the well's entire life. Over the course of a cycle (~3 years), a new portion of fluid is generated and accumulates in the relieved zone; the re-frac of the next cycle brings it into production together with the remaining undrained fluid.

Summed over N cycles, the slow contribution manifests as a systematic excess of actual cumulative production over the depletion forecast for a fixed S1 volume, and as anomalous well longevity (more than 30 years for the Eagle Ford vertical fleet), which is inexplicable within a single-charge reservoir model.

Fast rate recovery is redistribution through an existing network; long well life and reserves replenishment are the slow realization of S2 with the formation of a new network. This separate formulation is essential: it removes the kinetic objection (generation is not attributed to the fast contribution) while preserving the substantive core of the concept — S2 does work, but on the correct timescale.

5. Reinterpreting S2 and Tmax. Multi-parameter maturity control

5.1. S2 as an operational reserves category

In the shale paradigm, S2 is proposed to be entered into the reserves balance as an unrealized recoverable potential engaged by cyclic development. The tool for its quantitative estimation is the method of **potential balance** [1, 2, 6, 8]:

$$Q = S \cdot h \cdot \rho \cdot TOC \cdot HCyield \text{ (Table 2)},$$

Table 2.

Parameters of the potential balance method (formula Q = S · h · ρ · TOC · HCyield)

Parameter	Symbol	Units	Typical range / notes
Drained zone area	S	m²	Determined by SRV geometry
Effective thickness	h	m	Thickness of the source rock interval engaged in development through fracturing
Rock density	ρ	t/m³	2.4–2.8
Total organic carbon content	TOC	fraction	1–15
Specific hydrocarbon yield	HC_{yield}	m³/t	0.1 (oil) — 1 (dry gas)

The fraction of S2 realized over the project's life cycle is a calibratable parameter that can be measured experimentally.

5.2. Tmax: reinterpretation, not rejection

The Tmax parameter is retained, but its reading is corrected. First, classical Tmax maturity thresholds [14] are calibrated on conventional source rocks and do not account for systematic effects specific to a shale reservoir: the influence of heavy hydrocarbons and bitumoids retained in the pores on the shape of the S2 peak, matrix retention, and the effect of the S2 magnitude itself on the position of the maximum.

Second — and this is critical — within the logic of this work, Tmax of a formation under AHFP characterizes an **arrested, not maximally attainable,** state of the system: the formation may be "younger" in Tmax terms than its thermal history would suggest, precisely because pressure has arrested the realization of its potential. This yields a prospecting criterion: **low Tmax combined with high S2 in an AHFP zone** is not a sign of poor prospectivity but a signature of a charged, arrested reactor — that is, a priority target for cyclic development.

5.3. Maturity: from a single parameter to an ensemble

Direct Ro measurement is traditionally regarded as the maturity benchmark. However, a limitation must be honestly acknowledged: the target formations (Eagle Ford, Pimienta, Vaca Muerta, Domanik) are composed of marine Type II kerogen, in which vitrinite as a higher-plant maceral is rare or absent, and isolated grains are often reworked. Relying on Ro alone is methodologically vulnerable in these sections.

The solution is multi-parameter maturity control: Ro where vitrinite is representative; bitumen-reflectance equivalents recalculated via established correlations [26]; Tmax with the corrections of Section 5.2; molecular indicators (biomarker ratios, the methylphenanthrene index); pyrolysis kinetic spectra. Consistency across the ensemble of indicators is a reliability criterion; **divergence between Tmax and independent indicators within an AHFP zone** is a diagnostic sign of baric retardation — that is, **a standalone prospecting criterion**.

6. Mechanism of cyclic development (36+3)·N

Cycle scheme: 36 months of the working period + 3 weeks for near-wellbore zone stabilization ahead of re-fracturing and cleanup; N is the number of cycles, determined by the economic limit and the rate of response degradation. The physical content of the cycle (Table 3, Figure 5).

Table 3.

Physical content of the cyclic development cycle

Phase	Duration	Key processes
Production	36 months	Depression propagates through the SRV; the S1 component is produced; pressure in the affected volume drops below AHFP (significant-response threshold ~5 MPa); the baric brake is lifted in the relieved zone, and slow S2 conversion starts, accompanied by microfracturing. By 18–24 months the proppant degrades, the far-field fractures lose conductivity, and the drainage radius contracts — but the relieved far zone continues to generate and accumulate fluid.
Stabilization before re-fracturing	3 weeks	Preparation of the near-wellbore zone ahead of repeat hydraulic fracturing.
Repeat fracturing (re-frac)	1–2 days	Connection between the wellbore and the far zone is restored, where by this point the following have accumulated: (a) previously undrained existing fluid and (b) a new portion of hydrocarbons converted from S2. The rate recovers to 70–100% of the initial value.
Repeat × N	N cycles	Each cycle is a partial recharge of the system, not a "squeeze" of a diminishing remainder. Cumulative production grows faster than in a fixed-volume depletion model; the limiting N is set by economics and the actual dynamics of ΔS2.

Individual observations consistent with the scheme: productivity recovery after extended shut-ins at well 114 of the Salym field [1, 8, 9] (Figure 6), and long-term operation of the Eagle Ford vertical fleet [1, 8, 10] (Figure 7) beyond forecast depletion horizons.

Cycle parameters are not universal but are determined by the project's objective function. For onshore development, where the optimum is shifted toward maximizing the oil recovery factor per well, the (36 months + 3 weeks)·N scheme is justified. For offshore projects, where the principal asset is not the well but the platform, and downtime costs are disproportionately higher, the optimum shifts toward more aggressive production — up to a (24 months + 2 weeks)·N scheme with a shortened drainage-point exploitation period.

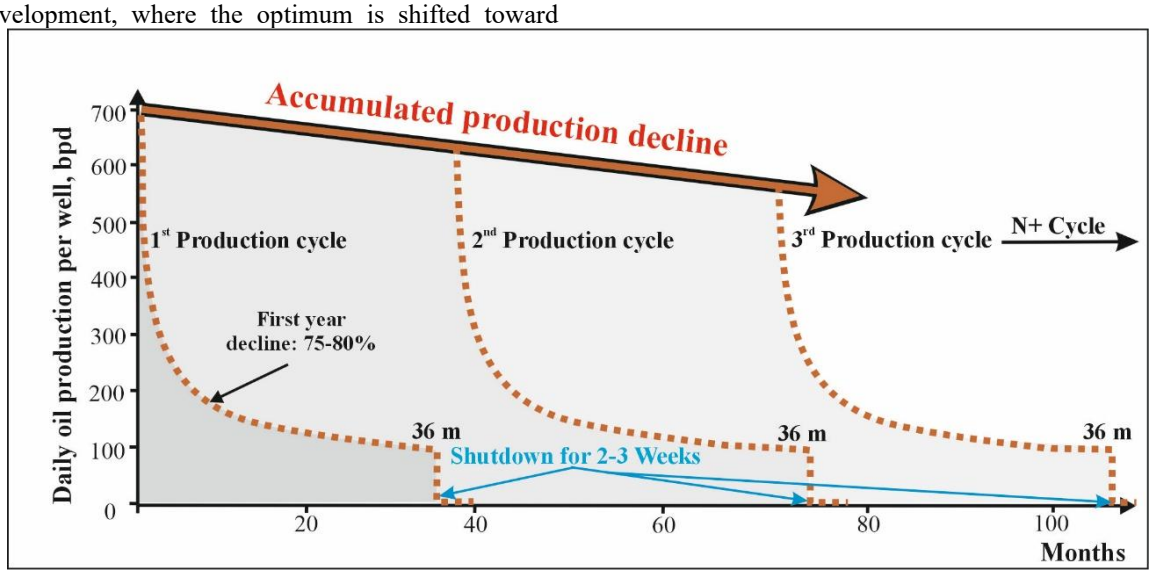


Figure 5. Principal scheme of shale formation cyclic development

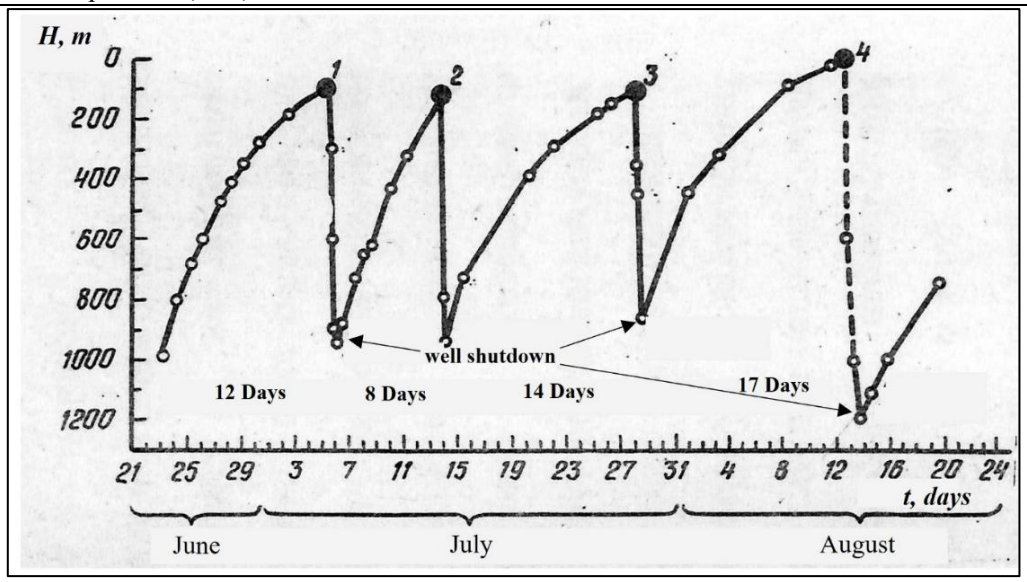


Figure 6. Periodic production history plot for well 114, Salym field [1, 8, 9]

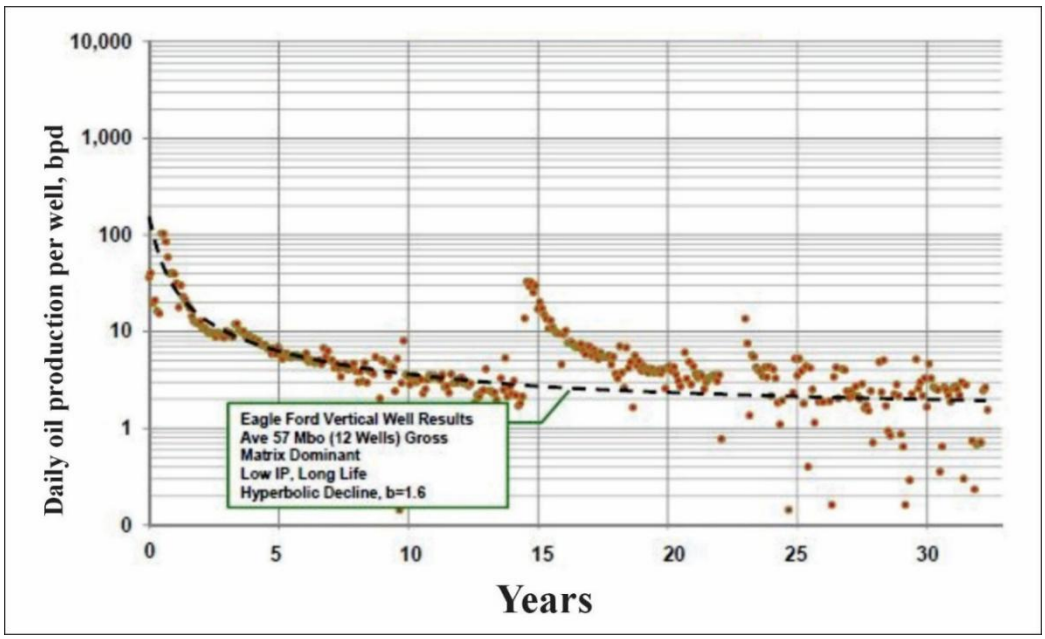


Figure 7. Chart summarizing results of studies conducted by EOG. Rate recovery of vertical(!) wells under periodic workovers and fracturing [1, 10]

7. Experiment: direct ΔS2 measurement

The strength of the concept lies in the fact that it allows direct quantitative verification by field means. An experiment is proposed comparing S2 before and after drainage by hydraulic fracturing.

Baseline profile (pilot well). The experiment relies on the natural sequence of drilling operations and requires no additional drilling for the sake of the experiment itself. Before the horizontal section is drilled, a vertical pilot well is drilled at the site: core is taken across the entire target interval, and a baseline Rock-Eval profile (S1, S2, Tmax, TOC) is obtained with a dense depth spacing. The horizontal section is then drilled from this same pilot, and the well is placed on cyclic development.

Repeat profile (offset vertical well). After 2–3 re-frac cycles (6.5–10 years), a new vertical well is drilled roughly 100 m from the horizontal section — at this distance it is guaranteed to fall within the fracturing drainage zone (SRV) of the horizontal well. Core is taken again at the same stratigraphic level and the same Rock-Eval profile is obtained. S2 in the baseline and repeat profiles is compared interval by interval, level to level, normalized to TOC.

The expected effect size is not postulated arbitrarily: it is calculated for the specific well from the potential balance — from cumulative production in excess of the S1 depletion forecast, converted into an equivalent S2 deficit over the drained-zone volume — and is compared against a detection threshold determined by Rock-Eval reproducibility. If the calculated ΔS2 falls below the detection threshold, the design is adjusted (more cycles, a zone with denser sampling).

The concept's separate architecture allows it to be tested not only through chemistry but also through microstructure — which substantially increases the experiment's resolving power. If S2 realization is indeed occurring and is accompanied by volumetric expansion,

then the repeat profile should show not only an S2 deficit but also an increase in secondary organic-matter porosity and microfracture density relative to the baseline profile.

Accordingly, a microstructural suite is performed on core from both wells — the baseline pilot and the repeat offset vertical — comprising FIB-SEM and X-ray microtomography with quantitative assessment of OM porosity and microfracture density associated with kerogen inclusions. Agreement between the two independent signals — S2 depletion (chemistry) and increased microfracturing (mechanics) — is the strongest confirmation of the concept available by field means. Supplementary measurements that further increase resolving power are also carried out: the shift in Tmax and in the production index PI = S1/(S1+S2) in the repeat profile; gas isotope composition across cycles (signatures of "freshly generated" gas); and kerogen kinetic spectra before and after.

The experiment does not affect project economics, given horizontal-well payback within 11–14 months of operation and an IRR of cyclic development exceeding 50–70% on average.

8. Economic rationale

Base case: a well after primary fracturing, with a characteristic decline profile, depletes its economically recoverable volume in ~3 years. The cyclic scheme adds N periods with rate recovery to 70–100% of the initial value after each re-frac, at incremental capital costs equal to the cost of re-fracturing and cleanup — without new drilling, pad construction, or infrastructure, which are already amortized within the base period. According to the authors' estimates, cumulative production per well increases 2–5-fold relative to a single-frac scheme, and the IRR of the incremental cash flow from cycles exceeds 50–70% (it may exceed 100%) — a profile rare for the industry, of "returns rising at almost zero incremental CAPEX" (Table 4, Figure 8). The potential balance method gives an upper-bound estimate of the engageable resource and allows ranking the well fleet by priority for conversion to cyclic development: the priority candidates are **AHFP zones with high S2 and "lagging" Tmax**.

Table 4.

Key economic indicators of cyclic development

Indicator	Value
Increase in cumulative production per well	2–5× relative to a single-frac scheme
Rate recovery after re-frac	70–100% of the initial value
IRR of the incremental cash flow from cycles	On average > 50–70% across the fleet; > 100% for many wells
Horizontal well payback	11–14 months of operation
Incremental cycle capital costs	Limited to the cost of re-fracturing and cleanup — no drilling, pad construction, or infrastructure

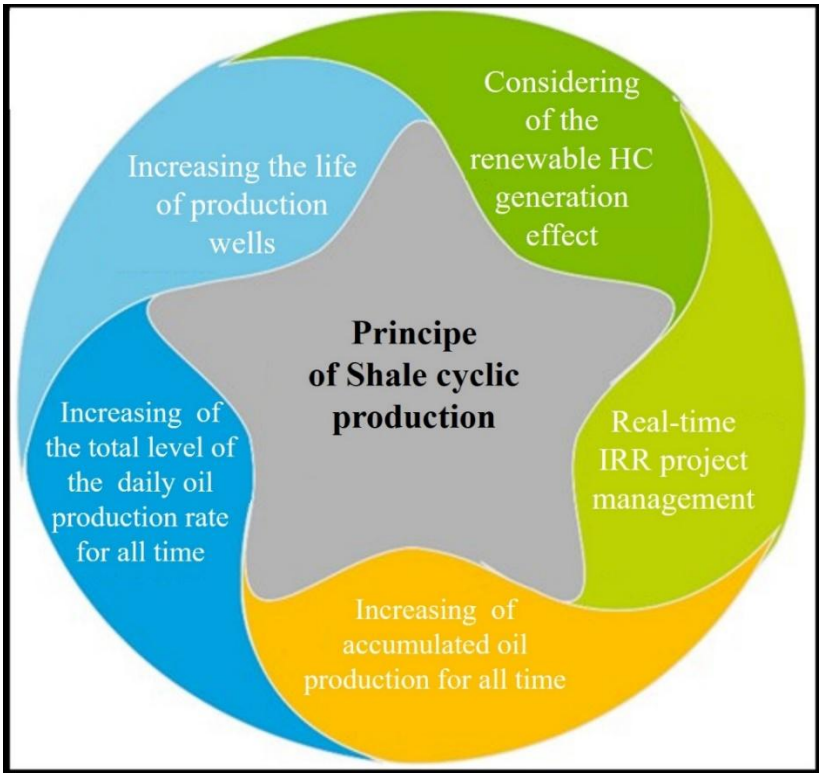

Considering of the renewable HC generation effect
Increasing the life of production wells
Principe of Shale cyclic production
Increasing of the total level of the daily oil production rate for all time
Real-time IRR project management
Increasing of accumulated oil production for all time

Figure 8. Economic basis of cyclic development of shale formations.

9. Responses to anticipated objections

Objection 1 (kinetic): Arrhenius kinetics at 100–130 °C will not produce noticeable S2 conversion over years

The answer is built into the architecture of the concept (Section 4): generation is not attributed to fast rate recovery — that is explained by redistribution. Generation is attributed to the slow, cumulative effect, for which low conversion rates are not a refutation but the working regime: even the first percentage points of S2 per year, integrated over the SRV volume and N cycles, yield production-significant volumes (verifiable by a potential-balance calculation for a specific well). Additionally, kinetic constants obtained by pyrolysis at atmospheric pressure do not account for the baric factor; the pressure-retardation literature [16–18] indicates that relieving excess pressure shifts the effective kinetics toward acceleration relative to the arrested state. Ultimately, the question is settled by the experiment.

Objection 2 (filtration-based): at nanodarcy permeability, newly generated hydrocarbons will not reach the wellbore

The objection refutes not the proposed concept but the classical model of inflow from the matrix, which is equally untenable at nanodarcy permeability. Transport in the proposed model is not diffusive but volumetric-baric: kerogen realization is accompanied by a volume increase that itself ruptures the matrix — a mechanism documented on Maykop shale core as hydraulic fracturing via a cavitation bubble collapsing with a hydraulic shock [5], consistent with independent estimates of the effect size and with laboratory/numerical modeling [20–24]; re-fracturing then connects the resulting micro-deposits to the wellbore. The mechanism is not postulated but rests on an experimentally observed and numerically reproduced phenomenon; what is new is the assertion of its reproducibility under cyclic pressure relief.

Objection 3: rate recovery and longevity are explainable without generation — by redistribution and pressure support from adjacent zones

This is true for the fast contribution. For the slow contribution, direct measurement provides confirmation: generation reduces S2 and increases OM porosity.

Objection 4: Ro is the maturity standard; abandoning it in favor of a reinterpreted Tmax is arbitrary

There is no abandonment: Ro retains priority wherever vitrinite is representative. But in the marine Type II kerogen of the target formations vitrinite is rare, and reliance on Ro alone is more vulnerable than the proposed ensemble of indicators. Reinterpreting Tmax is not a replacement for Ro but the extraction of additional information: divergence between Tmax and independent maturity indicators within an AHFP zone diagnoses baric retardation.

Objection 5: the Le Chatelier principle does not apply to irreversible cracking

The thermodynamic formulation is an interpretive framework; the evidentiary basis for the effect is empirical: pressure retardation has been reproduced experimentally and documented in sedimentary basins [16–18]. For the practical conclusions of this paper, it is sufficient to establish the empirical fact that **excess pressure retards the realization of potential, and depression removes that retardation**, regardless of whether the mechanism is described by equilibrium thermodynamics, a pressure correction to kinetics, or a combination of both.

10. Environmental rationale

The cyclic development scheme creates a qualitative shift not only in economics but also in the logic of field development. Current shale-drilling practice is geared toward the rapid rollout of the maximum number of well pads: since a single fracture network degrades within 1.5–2 years, and a well without re-fracturing depletes its economically recoverable volume within ~3 years, the only way to build and sustain a production plateau is to continuously bring new pads online, offsetting the decline of older ones. This sets a high pace of surface operations throughout the entire period of active field development.

The cyclic scheme removes this constraint at its root. Because every re-frac cycle delivers cumulative production 2–5 times higher than a single-frac scheme, at capital costs limited to the cost of the re-frac itself — that is, without new drilling, pad construction, or associated infrastructure — re-engaging the already-drilled well stock becomes a more capital-efficient use of investment than drilling a new pad. Under comparable (and, given the cumulative IRR effect, often higher) economics, the investment logic shifts: capital is better directed toward re-fracturing existing pads than toward drilling new ones. The practical consequence is that drilling of new pads can be deferred by 15–20 years — the time needed to fully exhaust the cyclic potential on pads already brought online — without any loss in production pace or any deterioration in project economics.

The environmental effect comprises several components:

- Reduction in one-time and cumulative land disturbance. The number of simultaneously operated well pads and their associated access roads, flowlines, and water infrastructure does not grow in proportion to the production rate but remains several times lower — the same volume of recovered hydrocarbons is achieved at a substantially smaller number of surface-disturbance points.
- Sparser well spacing as a systemic effect. With a partially replenishable reserves base, the optimal well spacing shifts toward sparser grids — fewer wells per unit area and more cycles per well. This directly reduces the project's total surface footprint per barrel of cumulative production.
- A load profile on infrastructure and logistics that is spread out over time. Peak intensity of equipment mobilization, water consumption for fracturing, and drilling-waste handling is not concentrated within a narrow window of intensive drilling but is spread over decades through periodic, localized re-frac operations at already-established pads — reducing the peak anthropogenic load on the local ecosystem and the social infrastructure of the operating area.
- Fuller use of impact already incurred. A well pad, access road, and pipeline tie-in, once built and having already caused land disturbance, continue to deliver

production over a substantially longer period, which increases the return on a unit of impact already incurred rather than requiring new impact.

A quantitative assessment of the effect (in terms of reduced disturbed land, mobilization emissions, and the volume of water used and disposed of per barrel of cumulative production) requires a separate calculation for a specific field and lies beyond the scope of this work; however, the direction of the effect itself — reproducible rather than one-time realization of potential on the existing well stock, instead of extensive drilling — appears significant in its own right, independent of the final numerical expression.

Table 5.

Traditional approach versus cyclic development: comparison of environmental profile

Criterion	Traditional approach (extensive drilling)	Cyclic development
Pace of new pad rollout	Continuous throughout the entire period of active development — new pads are brought online continuously to offset declining production from older ones	Deferred by 15–20 years — until the cyclic potential of already-online pads is fully exhausted
Cumulative surface footprint	Grows in proportion to production rate: the number of pads, roads, and pipelines increases with the volume of hydrocarbons recovered	Several times lower per barrel of cumulative production — the same output from substantially fewer surface-disturbance points
Well spacing density	Dense — the optimum is shifted toward the maximum number of wells per unit area	Sparse — fewer wells per unit area, more re-frac cycles per well
Peak logistical and infrastructure load	Concentrated within a narrow window of intensive drilling	Spread over decades through localized, periodic re-frac operations
Use of impact already incurred	One-time — the pad and infrastructure deliver production for a limited period (~3 years for a well without re-fracturing)	Repeated — the same pad and infrastructure operate for decades (anomalous well longevity)
Horizon for full field drill-out	Determined by drilling pace — typically minimized in time	Stretched over 15–20 years or more, synchronized with re-frac cycles

Conclusions

1. The classical interpretation of Rock-Eval pyrolysis is calibrated on conventional systems and systematically underestimates the development significance of S2 in shale formations, where the source and the reservoir coincide.

2. S2 is proposed to be interpreted as an unrealized hydrocarbon potential — an operational category of replenishable reserves, quantifiable by the potential balance method.

3. Formation pressure acts as a thermodynamic regulator of generation (the baric brake): AHFP arrests the realization of potential (the pressure-retardation empirics), and production depression resumes it.

4. Transport of newly generated hydrocarbons within the nanopermeable matrix is not diffusive: kerogen realization produces a 10–20% volumetric expansion that drives generation-induced microfracturing ("porous chocolate"). The hydrocarbons create their own migration pathways; the function of hydraulic fracturing is to connect the resulting micro-deposits to the wellbore (overcoming nano-immobility).

5. The novelty lies in asserting the reproducibility of this process: generation-induced microfracturing is not a single paleo-event of the maturation epoch but a controllable process re-triggered by every pressure-relief cycle.

6. Post-refrac rate recovery consists of a fast contribution (redistribution and desorption of existing fluids) and a slow, cumulative one (S2 realization, 2–3 years over N cycles). The separate formulation removes the kinetic objection and correctly positions the contribution of generation.

7. The concept is fully testable: an experiment is proposed with direct ΔS2 measurement (baseline profile of the pilot well before the cycle vs. repeat profile of the offset vertical well after N drainage cycles) and independent microstructural control (FIB-SEM, micro-CT), with a significance threshold set through the variance of lateral analogues.

8. The cyclic scheme (36 months + 3 weeks)·N implements the concept technologically: cumulative-production growth of 2–5× per well at CAPEX limited to re-fracturing, with an average incremental-flow IRR > 50–70% (for many wells IRR may exceed 100%).

The proposed concept does not reject but generalizes the tenets of petroleum geochemistry formulated in the 20th century with respect to hydrocarbon generation processes, pyrolysis, and the interpretation of the S2 parameter. The experience of active shale development accumulated in the 21st century provides an empirical basis unavailable to the method's founders: for the first time in the industry's history, the object of commercial development has become the generation kitchen itself, rather than a migration product. This makes it possible to refine and extend the core tenets of the sedimentary-migration theory of hydrocarbon origin - in particular, to move generation from the category of completed geological events into the category of managed processes taking place in the producing formation here and now. The practical value of this refinement extends beyond shale topics and is of interest to the oil and gas industry as a whole. Implementing cyclic development for carbonate shales (Pimienta, Permian, Eagle Ford, Vaca Muerte, Domanik, etc) will

substantially extend their productive life, increase cumulative production 2–5×, and achieve exceptional economic performance.

References

1. Бочкарев В.А. Концепция освоения сланцевых углеводородов – М.: ДПК Пресс, 2021. – 332 с. https://doi.org/10.25996/6492.2023.86.41.001.

2. Бочкарев В.А., Бочкарев А.В. Восполняемые запасы залежей углеводородов. – М.: ОАО «ВНИИОЭНГ». – 2017. – 276 с. https://doi.org/10.61726/3912.2024.76.35.001

3. Бочкарев А.В., Бочкарев В.А. Катагенез и прогноз нефтегазоносности недр // ВНИИОЭНГ. - 2006. - 321 с. https://doi.org/10.61726/9424.2024.69.60.001

4. Bochkarev V.A., Bochkarev A.V. Faults and hydrocarbons potentials. M.: JSC “VNIIOENG”, 2017. – 392p. https://doi.org/10.25996/6284.2024.97.81.001

5. Бочкарев А.В., Евик В.Н. Природа аномальных явлений в битуминозных глинах нижнего майкопа Центрального Предкавказья. – Литология и полезные ископаемые. - №1. – 1990. – С.59-71.

6. Бочкарев В.А., Лашманов Е.В., Сучок С.Н. Геологическое строение, формирование коллекторской ёмкости и особенности разработки месторождений нефти и газа в глинистых сланцах формации Игл Форд на юге США // Геология, геофизика и разработка нефтяных и газовых месторождений. 2010. № 7. С. 22–29.

7. Бочкарев В.А., Качалова Е.А. Сланцевый углеводородный потенциал формации Игл Форд (мексиканский сектор), условия его реализации и освоения // Нефтепромысловое дело. 2015. № 2. С. 5–14.

8. Bochkarev V.A. The Hydrocarbon Shales Production concept // VIII International Scientific and Practical Conference «Questions. Hypotheses. Answers: Science XXI Century», May 14–15, 2024, Toronto, Canada. P. 33–47. https://doi.org/10.5281/zenodo.11246386

9. Bochkarev V. Principles and approaches of the geological and geophysical studies, Development and Production of HC shale. Danish Scientific Journal, 82, 2024. P. 9-20. https://doi.org/10.5281/zenodo.10899563

10. Bochkarev, V.A.; Krivobokov, D.M.; Bochkarev, Art.V. Gas potential of Eagle Ford/ Agua Nueva formation in Mexico. Annali d’Italia, Scientific journal of Italy, 65, 2025. P. 11-20. https://doi.org/10.5281/zenodo.15089357

11. Bochkarev, V.A.; Krivobokov, D.M.; Bochkarev, Art.V. Development of the Pimienta shale formation –the key to the growth of daily oil production in Mexico. Znanstvena Misel, Scientific journal of Slovenia, 105, 2025. P. 11-20. https://doi.org/10.5281/zenodo.17012660

12. Espitalié J., Laporte J.L., Madec M., Marquis F., Leplat P., Paulet J., Boutefeu A. Méthode rapide de caractérisation des roches mères, de leur potentiel pétrolier et de leur degré d'évolution // Revue de l'Institut Français du Pétrole. 1977. V. 32. P. 23–42.

13. Tissot B.P., Welte D.H. Petroleum Formation and Occurrence. 2nd ed. Berlin: Springer-Verlag, 1984. 699 p.

14. Peters K.E. Guidelines for evaluating petroleum source rock using programmed pyrolysis // AAPG Bulletin. 1986. V. 70. № 3. P. 318–329.

15. Osborne M.J., Swarbrick R.E. Mechanisms for generating overpressure in sedimentary basins: a reevaluation // AAPG Bulletin. 1997. V. 81. № 6. P. 1023–1041.

16. Price L.C., Wenger L.M. The influence of pressure on petroleum generation and maturation as suggested by aqueous pyrolysis // Organic Geochemistry. 1992. V. 19. № 1–3. P. 141–159.

17. Carr A.D. A vitrinite reflectance kinetic model incorporating overpressure retardation // Marine and Petroleum Geology. 1999. V. 16. № 4. P. 353–377.

18. Hao F., Zou H., Gong Z., Deng Y. Petroleum migration and accumulation in the Bozhong sub-basin, Bohai Bay basin, China // AAPG Bulletin. 2007.

19. Запивалов Н.П., Попов И.П. Флюидодинамические модели залежей нефти и газа. Новосибирск: Изд-во СО РАН, филиал «Гео», 1999.

20. Kobchenko M., Panahi H., Renard F., Dysthe D.K., Malthe-Sørenssen A., Mazzini A., Scheibert J., Jamtveit B., Meakin P. 4D imaging of fracturing in organic-rich shales during heating // Journal of Geophysical Research: Solid Earth. 2011. V. 116. B12201.

21. Pelet R., Tissot B. Nouvelles données sur les mécanismes de genèse et de migration du pétrole // Proc. 8th World Petroleum Congress. Moscow, 1971. V. 2. P. 121–136.

22. Vernik L. Hydrocarbon-generation-induced microcracking of source rocks // Geophysics. 1994. V. 59. № 4. P. 555–563.

23. Teixeira M.G., Donzé F.-V., Renard F., Panahi H., Papachristos E., Scholtès L. Microfracturing during primary migration in shales // Tectonophysics. 2017. V. 694. P. 268–279.

24. Panahi H., Kobchenko M., Meakin P., Dysthe D.K., Renard F. Fluid expulsion and microfracturing during the pyrolysis of an organic rich shale // Fuel. 2019. V. 235. P. 1–16.

25. Behar F., Vandenbroucke M., Tang Y., Marquis F., Espitalié J. Thermal cracking of kerogen in open and closed systems: determination of kinetic parameters and stoichiometric coefficients for oil and gas generation // Organic Geochemistry. 1997. V. 26. № 5–6. P. 321–339.

26. Jacob H. Classification, structure, genesis and practical importance of natural solid oil bitumen («migrabitumen») // International Journal of Coal Geology. 1989. V. 11. P. 65–79; Landis C.R., Castaño J.R. Maturation and bulk chemical properties of a suite of solid hydrocarbons // Organic Geochemistry. 1995. V. 22. P. 137–149.

MEDICAL SCIENCES

OPTICAL ARCHITECTURE AND THE BULK-VERSUS-SPATIAL CONFOUND IN HYBRID LED PHOTOPOLYMERIZATION OF NAIL COATINGS

Fedorov M.
Photonic Technologies JSC, Samara, Optical Systems Development Engineer
DOI: 10.5281/zenodo.22107212

ABSTRACT

Hybrid, multi-wavelength LED curing lamps are increasingly marketed and adopted on the premise that combining two emission peaks resolves the spectral mismatch between broad-spectrum photoinitiator systems and single-wavelength light sources, producing more complete photopolymerization of pigmented coating films. This article argues that the evidence supporting this premise measures the wrong variable. Published comparisons report bulk or surface-averaged degree of conversion, a metric that can improve even as the spatial distribution of cure across the illuminated area remains uneven, because averaging across a heterogeneous beam profile raises the mean without correcting the variance. Drawing on photopolymerization kinetics, beam-profile radiometry, and the material-specific attenuation behavior of pigmented UV-curable films, the analysis traces this confound to its source and shows that it accounts for a direct contradiction between two recent systematic reviews addressing the same underlying question in a structurally identical material system. A parallel finding from instrumented training research complicates the proposed fix further: correcting the measurable variable does not automatically correct the chemical outcome that variable was meant to track. The article concludes that optical architecture for uniform nail-coating photopolymerization requires spatially resolved verification instead of bulk-averaged conversion data, and identifies the specific engineering step current hybrid LED designs have not yet implemented: position-differentiated sensor feedback in place of single-point measurement.

Keywords: hybrid LED curing, degree of conversion, beam profile inhomogeneity, photopolymerization uniformity, nail coatings, optical architecture.

Introduction

A hybrid LED lamp emitting simultaneously at 365 and 405 nanometers exists to solve one specific chemical problem: no single wavelength efficiently excites every photoinitiator present in a pigmented nail gel, so a lamp tuned to only one peak leaves photoinitiators absorbing outside that band under-activated regardless of exposure duration. Combining two emission peaks addresses this mismatch at the level of chemistry. The physical construction of a multi-diode array rarely produces a spatially uniform beam, because the individual LED dies sit at discrete positions across the light guide instead of fusing into a single emitting surface, and the resulting irradiance map is neither radially symmetric nor homogeneous across the treatment area [9]. Spectral coverage and spatial uniformity are two distinct engineering problems, and a lamp can solve the first while leaving the second untouched.

A third complication sits underneath both problems and is easy to overlook. Price et al. [14] point out that a curing unit's stated irradiance value is usually an averaged reading taken at the light tip, not a measurement of what actually reaches the material at a given distance and angle; the manufacturer's number describes the source. This distinction matters for hybrid nail lamps specifically, because a nail plate is neither flat nor fixed at a constant distance from the emitting surface across its width, so the gap between stated tip irradiance and delivered irradiance at the cuticle margin, at the center, and at the free edge of the nail can differ by construction, before any question of photoinitiator chemistry enters the picture.

The literature comparing hybrid and single-wavelength curing units has evaluated success almost exclusively through the chemistry problem's proxy: bulk-averaged degree of conversion, measured at a single point or averaged across a specimen, in place of the spatial distribution of conversion across the film. This choice of metric carries a measurable consequence, because the benefit hybrid architectures show over single-wavelength units in bulk-averaged terms is itself thickness-dependent and disappears past a documented point. Rocha et al. [16] found that the conversion advantage of a polywave unit over a monowave unit was pronounced up to roughly 2 millimeters of pigmented layer thickness and negligible beyond 2.5 millimeters, because violet-wavelength light attenuates faster through a pigmented matrix than blue light does, so the added spectral coverage that helps at the surface contributes little at depth. A metric that reports only the average across such a gradient obscures where along that gradient the material falls below threshold, the spatial information a uniformity claim needs and a bulk average cannot supply.

Gorbacheva [6] models this attenuation for pigmented nail gel films using the Bouguer-Lambert relationship, showing that irradiance falling on a curved, variably pigmented nail plate decreases with depth as a function of layer thickness and pigment loading, not lamp specification alone. That model describes attenuation through the material once light has entered it. It says nothing about whether the light entering different regions of the treatment area was uniform to begin with, the question this article isolates and treats as separable from the depth-attenuation question the model already answers. The distinction the field has not made, between spectral-coverage success and beam-profile success, is the specific gap this article resolves, and the reference point for that resolution appears by name in the

discussion below: position-resolved beam-profile measurement, the radiometric method Michaud et al. [9] and Price et al. [15] already apply to characterize dental curing units and that no published evaluation of a nail-curing hybrid LED system has yet adopted.

Literature Review

The case for hybrid architectures rests on a body of work establishing that spectral mismatch between photoinitiator and light source measurably degrades cure efficiency, and that combining wavelengths corrects this at the chemistry level. De Oliveira et al. [4] tested this directly by curing experimental composites containing camphorquinone alone, TPO alone, or blends of the two under a polywave unit, mapping the resulting cure profile by position beneath the beam's blue-emitting and violet-emitting regions separately. Light transmittance in the 380 to 420 nanometer range attenuated with depth more sharply than transmittance in the 420 to 495 nanometer range, so violet-sensitive photoinitiator systems cured less completely at depth than blue-sensitive ones under the same lamp. Lara et al. [8] extended this finding by combining camphorquinone and TPO in a single formulation, and found that the combined system consumed more of its photoinitiator load than the camphorquinone-only system did, evidence that the combination increases the fraction of available photoinitiator that reacts; overall degree of conversion showed no significant difference between the combined and single-initiator systems, which sets a practical limit on the size of the chemistry-level benefit even as the consumption data confirm the mechanism. Rocha et al. [16] supplied the thickness boundary on this benefit through FT-NIR microscopy mapping conversion along a cross-section, the methodological choice that made the 2 to 2.5 millimeter ceiling visible in the first place: surface conversion stayed high across all tested thicknesses even as the ceiling emerged at depth, a pattern a single-point surface measurement would have missed entirely. Oliveira et al. [12] pushed this same cross-sectional method one step further by varying the light-guide tip geometry itself, and found that the curing profile produced by a multiwave LED changed with the physical shape and emission pattern of the tip, which means the beam's spatial signature is not a fixed property of "polywave" as a category but a variable that depends on the specific tip design tested. This group of sources establishes something specific: the chemistry-level case for hybrid emission is real, explained by wavelength-dependent attenuation through pigmented material, and bounded by a thickness threshold and a tip-geometry dependency that became visible only once researchers measured conversion at more than one point and under more than one physical configuration.

Rocha et al.'s [16] cross-sectional method revealed a thickness-dependent boundary invisible to single-point measurement, and a second, independent literature asks whether the same logic applies laterally, across the treatment area. The answer complicates the case for hybrid systems in a way the first group of sources does not anticipate. Michaud et al. [9] characterized the irradiance output of several commercial dental light-curing units with a laser beam profiler and found that emission across the light-guide face was neither radially symmetric nor uniform, a geometric consequence of individual LED dies occupying fixed, discrete positions on the light guide instead of emitting as a single fused surface. Price et al. [13] found that this problem compounds for polywave units specifically: the violet-emitting and blue-emitting diodes within a single polywave array are physically distinct components occupying different positions on the light guide, so the two wavelengths are chemically complementary and spatially offset from one another, meaning a location on the treatment surface can receive strong blue irradiance and weak violet irradiance, or the reverse, depending on its position relative to each diode. Price et al. [15] quantified how this offset interacts with working distance: both irradiance magnitude and beam homogeneity degraded as the distance between lamp and surface increased across the four commercial units tested, a finding that bears directly on a curved, multi-nail treatment area where every finger cannot occupy the same distance from a fixed lamp at once. Wang et al. [19] closed the loop between beam inhomogeneity and material outcome by measuring microhardness, a validated proxy for degree of conversion, at the center and at points 3 millimeters left and right of center on cured specimens across four thicknesses and six shades, and found microhardness declined at off-center positions under the same nominal cure cycle and the same average irradiance value applied to the whole specimen; irradiance measured after passing through the specimen fell with both increasing thickness and darker shade.

A third group of sources locates this confound specifically within the nail-coating material system, and in doing so exposes a further variable that neither group above accounts for: formulation itself, not only lamp geometry or layer thickness, changes how much light a given region of coating actually receives. Bednarczyk and Rożniakowski [1] measured degree of conversion in a pigmented UV/LED nail gel directly by real-time FT-IR under dual 365/405 nanometer irradiation and found that different photoinitiator systems reached adequate conversion at different exposure times, a formulation-dependent variable layered on top of the beam-profile and depth-dependent variables the dental literature documents; a hybrid lamp calibrated to one formulation's optimal exposure time is not calibrated to another's. Mieriņa et al. [10] confirmed that this formulation dependence characterizes the nail-polish category broadly, reviewing polymerization mechanisms across acrylate, epoxide, and thiol-based nail-curable systems. Gorbacheva [7] adds a further attenuation source specific to nail products: nanoparticle additives such as titanium dioxide and silica, incorporated into modern coatings for their UV-protective and mechanical-reinforcement properties, function by absorbing or blocking UV wavelengths, so the nanoparticle load that protects the cured film from UV degradation after application also competes with the photoinitiator for incident light during the cure itself, a material-side attenuation source that operates independently of the diode-array geometry documented in the second group above. De Groot and Rustemeyer [3] and Steunebrink et al. [17] establish why this chain of attenuation sources matters

beyond film integrity: 2-hydroxyethyl methacrylate, the monomer whose incomplete conversion the whole chain traces back to, is the most frequently reacting allergen in patch-tested nail technicians and consumers, so a spatial cure deficit a bulk-averaged conversion measurement would miss marks a location on the nail where sensitizing monomer stays available for skin contact. This third group closes the literature review by supplying the material-specific reason the spatial confound identified in dental research is independently present and independently consequential in nail coatings, beyond simple analogy.

Discussion

The tension this literature review exposes resolves a disagreement that looks, on its surface, like a methodological dispute between two research teams and is in fact a measurement-design artifact with a specific, nameable fix. Fernández Godoy et al. [5] synthesized 26 in vitro studies and concluded that polywave units achieve significantly higher degree of conversion than monowave units, reporting a range of 83.7 to 92 percent against 70 to 81 percent. Thomaidis et al. [18], synthesizing a comparable body of bulk-fill composite studies at depths up to 4 millimeters, concluded that polywave units show no significant advantage over adequately powered monowave units at those depths. Fernández Godoy et al.'s [5] synthesis draws predominantly on studies using bulk or near-surface conversion measurements of the kind de Oliveira et al. [4] and the earlier group of sources here document, where hybrid emission's spectral-coverage advantage is real and measurable. Thomaidis et al.'s [18] synthesis draws on bulk-fill depth-of-cure studies operating past the 2.5-millimeter ceiling Rocha et al. [16] identified, where that same advantage has already attenuated to insignificance. Restated this way, the two reviews report results from two different regions of the same thickness gradient, and the field's failure to hold that gradient constant across studies, not a flaw in either review's methodology, produced the appearance of contradiction. This is the article's specific contribution: a demonstration that the contradiction dissolves once the compared studies are sorted by which region of the depth gradient they actually measured, a sorting the field has not performed and that this article's grouping of sources in the literature review above makes possible for the first time.

Gorbacheva's [6] sensor-feedback architecture, in which a microcontroller adjusts output to match a photoinitiator's absorption band and compensates for measured attenuation through a modeled layer thickness, addresses the chemistry-level and depth-attenuation problems this literature review's first and third groups document. The evidence from Michaud et al. [9] and Price et al. [13, 15] shows that a single-point or averaged sensor reading cannot detect the lateral beam-profile inhomogeneity those sources measure directly with a beam profiler instead of inferring from an aggregate output value. A sensor positioned to verify total emitted dose confirms that the lamp delivers adequate energy in aggregate; it does not confirm that the energy reaching the ring finger's cuticle margin matches the energy reaching the thumb's center, because those two locations sit at different points on the diode array's inherently uneven emission map regardless of how accurately the aggregate dose is calibrated. This qualification applies directly to the client's own architecture, not to a competing lamp design: the sensor-feedback mechanism Gorbacheva [6] describes solves the problem it was built to solve, spectral and depth calibration, and leaves untouched the spatially resolved problem Wang et al. [19] demonstrate empirically by measuring microhardness decline at off-center positions under nominally adequate average irradiance.

A finding from an adjacent training-evaluation literature places a further limit on how much this article's proposed fix can promise, and it needs to be stated directly. Burrer et al. [2] tested whether correcting the procedural variable under an operator's control, delivered irradiance, would also correct degree of conversion, using a patient simulator with real-time irradiance feedback for undergraduate dental students. Individualized, instrumented training produced a significant increase in delivered irradiance ($p < .001$) with markedly reduced variability across students, the kind of procedural improvement a position-resolved sensor architecture would also be built to achieve for a nail lamp operator. Degree of conversion itself, measured directly by FT-IR, showed no corresponding significant improvement. This result does not undermine the case for spatially resolved measurement built across the preceding sections; it establishes that measurement and correction are not the same step. A sensor array reporting irradiance at multiple points across the treatment area would diagnose where a nail lamp under-delivers, which is the capability current single-point systems lack entirely, but Burrer et al.'s [2] null result on conversion warns against assuming that diagnosis alone closes the gap between delivered light and converted monomer, since conversion also depends on the formulation- and thickness-dependent factors Bednarczyk and Rożniakowski [1] and Gorbacheva [7] document, none of which a corrected irradiance reading can compensate for on its own.

The engineering vocabulary to build a position-resolved system, as distinct from merely diagnosing its absence, already exists outside the dental and cosmetic literature. Moreno et al. [11] derive design equations for LED arrays intended to produce uniform near-field irradiance based on étendue conservation, the same non-imaging optics principle that governs how a source's angular and spatial extent limits how evenly its output can be redistributed by a downstream lens or reflector. Applied to a hybrid nail lamp, this principle sets a hard constraint the sensor-feedback question in the previous paragraph does not touch: no amount of downstream sensing or microcontroller compensation corrects for a diode array whose physical layout was never designed, at the optical level, to project a uniform combined field across a curved five-finger surface. Michaud et al.'s [9] and Price et al.'s [15] beam-profiler measurements describe the resulting non-uniformity in existing dental units; Moreno et al.'s [11] framework describes the redesign step that would prevent it at the source, an option that operates upstream of, and independently from, any sensor correction applied after the fact.

Conclusion

This analysis reframes what "uniform photopolymerization" should mean for hybrid LED nail-curing systems: not a bulk-averaged degree of conversion figure, which the existing literature already shows can rise while spatial deficits persist undetected, but a position-resolved conversion or proxy-hardness map verified with the same multi-point radiometric method dental curing-light research already applies to its lamps. Read this way, the apparent conflict between Fernández Godoy et al. [5] and Thomaidis et al. [18] stops being a disagreement to adjudicate and becomes a demonstration of exactly which region of the depth gradient each result belongs to, a distinction the field can now apply to new studies going forward instead of resolving retrospectively study by study. Burrer et al.'s [2] null result on conversion despite corrected irradiance adds a second, sharper conclusion: sensor-based diagnosis of spatial non-uniformity is a precondition for fixing the problem, and any evaluation protocol adopted for nail-curing lamps needs to measure conversion itself at multiple positions, not irradiance alone, if it is to avoid repeating the same confound this article has traced through the dental-composite literature. One question stays open: can a formulation-aware, spatially resolved sensor architecture, extending Gorbacheva's [6] single-point feedback model to multiple discrete measurement locations across the treatment area and recalibrating against the nanoparticle-dependent attenuation Gorbacheva [7] documents, fit within the size and cost constraints of a consumer or salon-grade nail lamp, or does spatial uniformity at this level of resolution require the fixed, factory-calibrated optical redesign Moreno et al.'s [11] etendue-based framework describes, applied at the diode-array stage before any sensor is added at all?

References

1. Bednarczyk, P., & Rożniakowski, K. (2025). The influence of the photoinitiating system on residual monomer contents and photopolymerization rate of a model pigmented UV/LED nail gel formulation. Coatings, 15(10), 1125. https://doi.org/10.3390/coatings15101125

2. Burrer, P., Par, M., Attin, T., & Tauböck, T. T. (2021). Real-time feedback of the applied light-curing technique and its impact on degree of conversion of composite restorations — A study with undergraduate dental students. Journal of Personalized Medicine, 11(10), 1012. https://doi.org/10.3390/jpm11101012

3. de Groot, A. C., & Rustemeyer, T. (2023). 2-Hydroxyethyl methacrylate (HEMA): A clinical review of contact allergy and allergic contact dermatitis. Part 1. Introduction, epidemiology, case series and case reports. Contact Dermatitis, 89(6), 401–433. https://doi.org/10.1111/cod.14405

4. de Oliveira, D. C. R. S., Rocha, M. G., Correa, I. C., Correr, A. B., Ferracane, J. L., & Sinhoreti, M. A. C. (2016). The effect of combining photoinitiator systems on the color and curing profile of resin-based composites. Dental Materials, 32(10), 1209–1217. https://doi.org/10.1016/j.dental.2016.06.010

5. Fernández Godoy, E., Chaple Gil, A., Caviedes Thomas, R., Bersezio Miranda, C., Martín Casielles, J., Rodríguez Martínez, G., & Angel Aguirre, P. (2025). Efficiency and limitations of polywave light-curing units in restorative dentistry: A systematic review. Clinical Oral Investigations, 29, 384. https://doi.org/10.1007/s00784-025-06457-4

6. Gorbacheva, Iu. (2025). Интеллектуальные UV/LED-системы для безопасного моделирования ногтей [Intelligent UV/LED systems for safe nail modeling]. LAP LAMBERT Academic Publishing.

7. Gorbacheva, Y. Y. (2025). Intelligent UV/LED systems for safe nail modeling [Electronic resource]. LAP LAMBERT Academic Publishing House.

8. Lara, L., Rocha, M. G., de Menezes, L. R., Correr, A. B., Sinhoreti, M. A. C., & Oliveira, D. (2021). Effect of combining photoinitiators on cure efficiency of dental resin-based composites. Journal of Applied Oral Science, 29, e20200467. https://doi.org/10.1590/1678-7757-2020-0467

9. Michaud, P. L., Price, R. B., Labrie, D., Rueggeberg, F. A., & Sullivan, B. (2014). Localised irradiance distribution found in dental light curing units. Journal of Dentistry, 42(2), 129–139. https://doi.org/10.1016/j.jdent.2013.11.014

10. Mieriņa, I., Grigale-Sorocina, Z., & Birks, I. (2025). The chemistry behind UV-curable nail polishes. Polymers, 17(9), 1166. https://doi.org/10.3390/polym17091166

11. Moreno, I., Avendaño-Alejo, M., & Tzonchev, R. I. (2006). Designing light-emitting diode arrays for uniform near-field irradiance. Applied Optics, 45(10), 2265–2272. https://doi.org/10.1364/AO.45.002265

12. Oliveira, D., Rocha, M. G., Correr, A. B., Ferracane, J. L., & Sinhoreti, M. (2019). Effect of beam profiles from different light emission tip types of multiwave light-emitting diodes on the curing profile of resin-based composites. Operative Dentistry, 44(4), 365–378.

13. Price, R. B. D., Labrie, F. A., Rueggeberg, F., & Felix, C. M. (2010). Irradiance differences in the violet (405 nm) and blue (460 nm) spectral ranges among dental light-curing units. Journal of Esthetic and Restorative Dentistry, 22(6), 363–377.

14. Price, R. B., Ferracane, J. L., & Shortall, A. C. (2015). Light-curing units: A review of what we need to know. Journal of Dental Research, 94(9), 1179–1186. https://doi.org/10.1177/0022034515594786

15. Price, R. B., Labrie, D., Whalen, J. M., & Felix, C. M. (2011). Effect of distance on irradiance and beam homogeneity from 4 light-emitting diode curing units. Journal of the Canadian Dental Association, 77, b9.

16. Rocha, M. G., de Oliveira, D. C. R. S., Correa, I. C., Correr-Sobrinho, L., Sinhoreti, M. A. C., Ferracane, J. L., & Correr, A. B. (2017). Light-emitting diode beam profile and spectral output influence on the degree of conversion of bulk fill composites. Operative Dentistry, 42(4), 418–427. https://doi.org/10.2341/16-164-L

17. Steunebrink, I. M., de Groot, A., & Rustemeyer, T. (2024). Contact allergy to acrylate-containing nail cosmetics: A retrospective 8-year study. Contact Dermatitis, 90(3), 262–265. https://doi.org/10.1111/cod.14475

18. Thomaidis, S., Masouras, K., & Papazoglou, E. (2026). Bulkfill resin composite polymerization efficiency by monowave vs. polywave light curing units: A systematic review of in vitro studies. Applied Sciences, 16(1), 346. https://doi.org/10.3390/app16010346

19. Wang, L., Ku, H., Kim, D., & Park, S.-H. (2024). Effect of shade and thickness on the microhardness of resin-based composite specimens at different points considering curing light beam's inhomogeneity. BMC Oral Health, 24(1), 1–8. https://doi.org/10.1186/s12903-024-04647-2

PHILOLOGICAL SCIENCES

THE IMPACT OF MULTILINGUALISM ON SOCIAL INTEGRATION

Allahverdiyeva A.
PhD., Ass.Prof.
Azerbaijan National Academy of Sciences(ANAS)
Nasimi Institute of Linguistics, Department of Theoretical Linguistics
Azerbaijan, Baku
https://orcid.org/0009-0000-6660-0357
DOI: 10.5281/zenodo.22107229

ABSTRACT

Multilingualism has become one of the defining characteristics of contemporary societies due to increased globalization, international migration, economic cooperation, and cultural exchange. The ability of individuals and communities to use more than one language influences not only communication processes but also social relationships, cultural identity, educational opportunities, and participation in public life. This article examines the relationship between multilingualism and social integration by exploring how linguistic diversity affects inclusion, intercultural understanding, and social cohesion.

The study is based on a qualitative analysis of existing academic literature in the fields of sociolinguistics, bilingual education, language policy, and migration studies. The research analyzes theoretical approaches and previous findings concerning the role of multilingual competence in creating connections between different social groups. The results indicate that multilingualism can positively contribute to social integration by reducing communication barriers, strengthening cultural awareness, and improving access to educational and professional opportunities. At the same time, the study highlights that linguistic inequality, discrimination against minority languages, and insufficient institutional support may limit the positive effects of multilingualism.

The article argues that multilingualism should be considered a valuable social resource rather than simply an individual communication skill. Effective language policies and inclusive educational systems are essential for transforming linguistic diversity into a factor that promotes equality, participation, and social development.

Keywords: multilingualism, social integration, linguistic diversity, bilingual education, intercultural communication, language policy, multicultural society.

Introduction

Language represents one of the most important foundations of human society. It enables individuals to exchange information, express cultural identities, establish relationships, and participate in social institutions. In the modern world, linguistic diversity has become increasingly visible due to globalization, migration, international education, and technological development. As a result, multilingualism has gained significant importance in academic research and social policy discussions.

Multilingualism refers to the ability of individuals or communities to use two or more languages for communication. It is not limited to knowledge of vocabulary and grammar but also includes the ability to function within different cultural and social contexts. According to Grosjean (2010, p. 13), multilingual individuals develop unique communicative abilities that allow them to interact with various linguistic communities while maintaining connections with different cultural environments. Therefore, multilingualism represents both a linguistic and a social phenomenon.

Social integration is the process through which individuals and groups become active members of society while developing relationships with others and participating in social, economic, and cultural activities. Language plays a central role in this process because communication is necessary for accessing education, employment, healthcare, government services, and social networks. Individuals who possess multilingual abilities may have greater opportunities to participate in diverse communities and establish meaningful connections across cultural boundaries.

In multicultural societies, language often becomes a key factor influencing the relationship between majority and minority groups. Immigrant communities, for example, frequently maintain their native languages while learning the dominant language of the host country. This multilingual reality creates both opportunities and challenges. On one hand, maintaining heritage languages can strengthen cultural identity and community belonging. On the other hand, insufficient knowledge of the dominant language may create barriers to social participation and economic mobility (Baker, 2011, p. 102).

Previous research has demonstrated that multilingualism can contribute positively to social cohesion when linguistic diversity is recognized and supported by institutions. García (2009, p. 45) emphasizes that bilingual and multilingual practices can provide learners with broader cultural perspectives and improve their ability to participate in diverse social environments. Similarly, May (2012, p. 78) argues that language diversity can become a resource for democratic participation when minority languages receive appropriate recognition.

One of the most important contributions of multilingualism to social integration is its ability to facilitate intercultural communication. People who speak multiple languages often develop greater awareness of cultural differences and similarities. Through language learning, individuals gain access to different ways of

thinking, social values, and cultural traditions. This interaction may reduce stereotypes and promote mutual understanding among different communities.

Multilingualism also has important implications for education. Educational systems that support students' linguistic backgrounds can improve academic achievement and encourage social inclusion. When students are encouraged to maintain their home languages while learning additional languages, they often develop stronger identities and greater confidence in multicultural environments. According to Cummins (2000, p. 37), recognizing students' linguistic resources can positively influence both educational outcomes and social development.

However, the relationship between multilingualism and integration is not always positive. In many societies, certain languages are considered more valuable or prestigious than others. Dominant languages often provide greater access to economic and political opportunities, while minority languages may be marginalized. This situation can create linguistic inequalities and pressure individuals to abandon their cultural and linguistic heritage. Skutnabb-Kangas (2000, p. 45) highlights that the suppression of minority languages can negatively affect cultural diversity and social equality.

Furthermore, migrants may experience difficulties when adapting to new linguistic environments. Limited proficiency in the dominant language can restrict access to employment, education, and public services. Therefore, successful social integration requires balanced language policies that promote acquisition of the dominant language while protecting linguistic diversity.

The increasing importance of multilingualism in modern societies has encouraged researchers to examine its broader social consequences. Studies in sociolinguistics, education, and migration research suggest that language is not merely a tool for communication but also a mechanism through which social identities and relationships are constructed.

The purpose of this article is to analyze the impact of multilingualism on social integration by examining its advantages, challenges, and social implications. The study focuses on how multilingual abilities influence communication, cultural participation, educational opportunities, and inclusion within multicultural societies.

Methods

This study employs a qualitative research methodology based on a comprehensive review and analysis of academic literature related to multilingualism and social integration. The purpose of the methodological approach is to examine existing theoretical perspectives and empirical findings concerning the relationship between language diversity and social participation.

The research is conducted through a qualitative content analysis of scholarly publications in the fields of sociolinguistics, applied linguistics, education, migration studies, and language policy. The selected sources include academic books, peer-reviewed journal articles, and research studies that examine multilingual practices in different social and cultural contexts. Particular attention is given to studies focusing on immigrant communities, multilingual education, minority language rights, and intercultural communication.

The literature was analyzed according to several main themes. The first theme concerns the role of multilingualism in facilitating communication and developing social relationships. The second theme examines the relationship between multilingual competence and access to educational and professional opportunities. The third theme focuses on the challenges associated with multilingual societies, including language inequality, discrimination, and institutional barriers.

A qualitative approach was selected because it allows for a deeper understanding of the social meanings and consequences of multilingual practices. Rather than measuring multilingualism only through numerical indicators, the study investigates how language abilities influence individuals' experiences, identities, and participation within society.

The research does not include primary data collection through interviews or surveys. Instead, it synthesizes findings from previous academic studies in order to provide a broader theoretical understanding of how multilingualism contributes to or challenges social integration. The reviewed literature represents different geographical and cultural contexts, allowing the study to identify common patterns in multilingual societies.

Results

The analysis of academic literature reveals that multilingualism has a significant influence on social integration. The findings demonstrate that multilingual abilities can support inclusion by improving communication, strengthening cultural relationships, and increasing participation in social institutions. However, the positive effects of multilingualism depend largely on social attitudes, educational systems, and language policies.

Multilingualism and Social Communication

One of the main findings of the study is that multilingualism improves communication between individuals from different linguistic backgrounds. Language is a fundamental instrument for establishing relationships, and knowledge of multiple languages enables people to interact with wider social networks.

Multilingual individuals often have the ability to communicate with both their own linguistic communities and members of other groups. This ability creates opportunities for cooperation and reduces social isolation. In multicultural environments, multilingual speakers frequently act as intermediaries who facilitate communication between communities.

For immigrant populations, multilingual competence can be particularly important. Migrants who maintain their native language while learning the language of the host society often develop stronger connections with both their heritage communities and the wider population. This dual linguistic ability allows them to participate in different social environments without completely losing their cultural identity.

Grosjean (2010, p. 85) argues that multilingual individuals should not be viewed as possessing separate language systems but as people who use their complete linguistic resources according to different social situa-

tions. This perspective demonstrates that multilingualism enables individuals to adapt to various communication contexts and develop flexible social identities.

Multilingualism and Cultural Understanding

The results also show that multilingualism contributes to intercultural awareness and cultural understanding. Learning another language usually involves learning about the history, traditions, values, and perspectives associated with that language. As a result, multilingual individuals may develop greater openness toward cultural differences.

In socially diverse communities, multilingual communication can create spaces for dialogue between groups. When people are able to communicate across cultural boundaries, misunderstandings may decrease and cooperation may increase. Language knowledge therefore contributes not only to communication but also to the development of social trust.

According to May (2012, p. 91), linguistic diversity can strengthen social cohesion when societies recognize different languages as valuable components of collective identity. Instead of viewing multiple languages as obstacles, inclusive societies can use linguistic diversity as a foundation for cultural exchange.

Multilingual education provides an important example of this process. Schools that support students' home languages while teaching additional languages encourage inclusion and respect for diversity. Such educational approaches allow students to maintain connections with their cultural backgrounds while developing the skills needed for participation in wider society.

Multilingualism and Educational Integration

Education represents one of the most important areas where multilingualism affects social integration. Students from multilingual backgrounds may experience educational advantages when their linguistic abilities are recognized and supported.

Research shows that maintaining a student's first language while acquiring additional languages can contribute to cognitive development, academic achievement, and confidence. Cummins (2000, p. 68) emphasizes that educational systems should consider students' existing linguistic knowledge as a resource rather than a limitation.

In contrast, educational environments that ignore or discourage students' home languages may negatively affect their sense of belonging. When minority languages are excluded from schools, students may feel that their cultural identities are undervalued. Therefore, inclusive language policies are essential for creating equal educational opportunities.

Multilingualism and Economic Participation

Another significant finding is the relationship between multilingual skills and economic opportunities. In contemporary global economies, language abilities are increasingly valuable in fields such as international business, tourism, diplomacy, education, and healthcare.

Multilingual employees can communicate with international partners, customers, and diverse communities more effectively. As a result, multilingual competence may increase professional opportunities and economic mobility.

However, the economic value of languages is not equally distributed. Some international languages provide greater career advantages, while minority languages may receive limited recognition. This difference demonstrates that social and economic systems influence how multilingualism is valued.

Challenges and Linguistic Inequality

Despite the benefits of multilingualism, the research identifies several challenges that may limit its positive impact on integration. One major challenge is linguistic inequality, which occurs when certain languages are considered superior or more useful than others.

Minority language speakers may face discrimination or pressure to adopt only the dominant language. Such situations can lead to language loss and weaken cultural connections. According to Skutnabb-Kangas (2000, p. 112), protecting linguistic rights is essential for maintaining cultural diversity and ensuring social equality.

Another challenge concerns migrants who have limited knowledge of the dominant language in their new country. Without adequate language support, they may experience difficulties accessing education, employment, and public services. Effective integration policies must therefore provide opportunities for language learning while respecting individuals' existing linguistic backgrounds.

Overall, the results demonstrate that multilingualism has the potential to strengthen social integration when supported by inclusive institutions and positive social attitudes. Language diversity becomes a social advantage when communities recognize and value different forms of linguistic expression.

Conclusion

Multilingualism represents an important social phenomenon that significantly influences the process of social integration in contemporary societies. The analysis conducted in this article demonstrates that the ability to communicate in multiple languages provides individuals and communities with valuable opportunities for interaction, cultural exchange, education, and participation in social life.

The results indicate that multilingualism contributes positively to social integration by reducing communication barriers and creating stronger connections between people from different linguistic and cultural backgrounds. Multilingual individuals often function as bridges between communities because they can facilitate communication and promote mutual understanding. Their linguistic abilities allow them to participate in multiple social networks while maintaining connections with their cultural heritage.

The study also highlights the importance of multilingualism in education and professional development. Educational systems that support linguistic diversity can improve students' academic experiences and strengthen their sense of belonging. Similarly, multilingual skills can provide economic advantages by increasing access to employment opportunities in increasingly globalized labor markets.

However, multilingualism does not automatically guarantee social integration. The social value of languages depends on political, cultural, and institutional

conditions. When minority languages are ignored or marginalized, linguistic diversity may become associated with inequality rather than inclusion. Therefore, language policies play a crucial role in determining whether multilingualism functions as a social resource or becomes a source of exclusion.

Inclusive approaches to language education and social policy are necessary to maximize the benefits of multilingualism. Governments and educational institutions should encourage the learning of additional languages while also protecting minority and heritage languages. Such approaches can support cultural diversity, strengthen social cohesion, and promote equal participation among different groups.

In conclusion, multilingualism should be understood as more than an individual ability to speak several languages. It is a social resource that contributes to communication, identity formation, cultural understanding, and community development. In increasingly diverse societies, recognizing and supporting multilingualism is essential for building inclusive and socially integrated communities.

References

1. Baker, C. (2011). *Foundations of Bilingual Education and Bilingualism* (5th ed.). Bristol: Multilingual Matters. 496 pp.

2. Cummins, J. (2000). *Language, Power and Pedagogy: Bilingual Children in the Crossfire*. Clevedon: Multilingual Matters. 320 pp.

3. García, O. (2009). *Bilingual Education in the 21st Century: A Global Perspective*. Malden, MA: Wiley-Blackwell. 496 pp.

4. Grosjean, F. (2010). *Bilingual: Life and Reality*. Cambridge, MA: Harvard University Press. 288 pp.

5. May, S. (2012). *Language and Minority Rights: Ethnicity, Nationalism and the Politics of Language* (2nd ed.). New York: Routledge. 384 pp.

6. Piller, I. (2016). *Linguistic Diversity and Social Justice: An Introduction to Applied Sociolinguistics*. Oxford: Oxford University Press. 224 pp.

7. Portes, A., & Rumbaut, R. G. (2014). *Immigrant America: A Portrait* (4th ed.). Berkeley: University of California Press. 472 pp.

8. Skutnabb-Kangas, T. (2000). *Linguistic Genocide in Education—or Worldwide Diversity and Human Rights?* Mahwah, NJ: Lawrence Erlbaum Associates. 824 pp.

9. Spolsky, B. (2004). *Language Policy*. Cambridge: Cambridge University Press. 256 pp.

10. Wei, L. (2018). *The Cambridge Handbook of Linguistic Multi-Competence*. Cambridge: Cambridge University Press. 550 pp.

SOCIAL SCIENCES

INTERNATIONAL EXPERIENCE IN THE IMPLEMENTATION OF STATE SOCIAL ORDERS FOR SOCIAL SERVICES BY NGOs

Aliyev U.
PhD in Sociology
Lecturer, Department of Humanities, Baku Business University
Orcid: 0000-0002-1615-9062
DOI: 10.5281/zenodo.22107257

ABSTRACT

In the article, the international experience of implementing the state social order in the field of social services by non-governmental organizations (NGOs) is analysed from both theoretical and practical perspectives. First of all, the principles governing the organization of the social services system in foreign countries, its legal regulation, the mechanisms for the provision of social services, as well as the main directions of the development of the social services institution in Western countries, are examined. Subsequently, the growing role of the third sector in the provision of social services in developed countries, the mechanisms of cooperation between NGOs and public authorities, their organizational forms and operational characteristics, as well as their advantages in implementing the state social order, are assessed. It is established that the types and scope of social services are not only closely related to the level of countries' economic development but also that the active participation of civil society institutions contributes significantly to improving the quality, accessibility, and efficiency of social service delivery. The article substantiates the significance for Azerbaijan of the social services convergence approach, the principles of social insurance, and the state–NGO partnership model applied in Western European countries. Taking into account international experience, the main principles of state support are identified, and proposals are put forward for strengthening the criteria of inclusiveness and social responsibility in the implementation of the state social order, particularly for the introduction of legal mechanisms promoting the employment and rehabilitation of persons with disabilities. Finally, the necessity of strengthening institutional cooperation among the state, NGOs, and society in ensuring the effective realization of the right to social services is substantiated.

Keywords: social service, state social order, non-governmental organizations (NGOs), third sector, state–NGO partnership, regulation of social services, international experience, civil society.

Introduction

The development of the concept of the social state, the strengthening of the social protection of the population, and the enhancement of the quality of social services stand out as priority directions of a state's social policy in the contemporary period. The active participation not only of state institutions but also of non-governmental organizations (NGOs) — the principal subjects of civil society — is of particular importance in achieving these objectives. Over recent decades, many countries around the world have established institutional foundations for cooperation between the state and non-state sectors in the provision of social services, and the state social order has become one of the most effective mechanisms of this cooperation.

The state social order mechanism creates the conditions for a competitive environment in the provision of social services, for a more flexible and targeted organization of services, for the strengthening of public oversight, and for a more efficient use of public funds. In particular, the innovative approaches of NGOs, their close ties with local communities, their volunteering potential, and their capacity to act according to individual needs in delivering social services to vulnerable population groups constitute their principal advantages in implementing state social orders. In this respect, the delegation of part of social service provision by the state to NGOs is regarded as one of the important means of developing the partnership model in social governance and of raising the efficiency of the social services system.

International experience shows that the implementation of state-commissioned social services by NGOs has taken shape on the basis of various legal and organizational models. The application of this mechanism in individual countries is conditioned by the model of public administration, the specific features of social policy, the level of development of the non-profit sector, and the principles of financing. At the same time, transparent competitive procedures, results-based financing, monitoring of service quality, and accountability constitute the main elements of the implementation of state social orders in most countries. Although the concrete mechanisms and institutional approaches differ from country to country, the general trend is directed toward expanding the partnership between the state and NGOs in the provision of social services. This model serves not only the more flexible, targeted, and efficient organization of social services but also the strengthening of public participation and the role of civil society in the implementation of social policy.

The purpose of the article is to analyse, in a comparative perspective, the international experience accumulated in the field of the implementation of state social orders for social services by NGOs; to identify the specific features of the legal and institutional models applied in various countries; to evaluate their strengths and weaknesses; and to substantiate the possibilities of applying this experience to the improvement of the social services system and the development of the state–NGO partnership in the Republic of Azerbaijan.

1. Features of the Regulation of the Social Service Sector in Foreign Countries

The expansion of the social service system that accompanied the economic development of Western countries has taken place along three directions.

First, the broadening of the risks covered by the provision of social services: whereas the social service system initially covered protection against only certain social risks, it subsequently began to encompass a wide range of them.

Second, the broadening of the range of persons under social protection: whereas the social service system initially covered only wage earners, it gradually began to cover other categories of persons in need of protection as well (for example, craftsmen, certain strata of the unemployed, persons undergoing rehabilitation, and recipients of social assistance).

Third, the growth in the quantity and the rise in the level of the types of social services provided: although the number of services offered was limited in earlier periods, the level of services provided has gradually risen along with the growth of states' economic potential.

Research conducted in economically developed countries in the field of social protection shows that the organization of the social service system rests mainly on two principles. According to the first principle, the state supports the activity of civil society institutions — in particular NGOs and other public associations — in the provision of social services and delegates part of its own powers to them, as well as to local self-government bodies. Under market conditions, this approach enables functional decentralization and allows for a more flexible and efficient organization of social services. According to the second principle, although various actors participate in the provision of social services, the principal responsibility for guaranteeing citizens' social rights and creating decent living conditions for them remains with the state. In other words, the transfer of part of social service provision to non-state entities does not remove the state's social obligations in this sphere; on the contrary, the state continues to act as the actor bearing the main responsibility for the formulation, financing, and outcomes of social policy [6; 7].

German legislation enshrines the following principles of social services: the insurance principle (financing of social services through social insurance contributions, and a correlation between the level of social insurance payments and the volume of social services provided); the provision principle (financing through taxes); and the guardianship principle (financing from the budget and targeting of individual needs) [11].

In the United Kingdom, the regulation of social services rests on the following principles:

- universality (social services extend to the entire population of the country);
- comprehensiveness (all social risk situations are covered);
- unity of administration (social services are administered by a unified public service);
- uniformity of the services provided (services must be the same regardless of supplements or of a person's circumstances, and must not depend on the amount of income) [9, pp. 938–959].

In the United States, the state social service system — which directly incorporates social work as one of its components — emerged considerably later than in the European countries. For a long time, social service issues did not receive adequate attention in the United States, and state social policy was directed mainly toward pension provision and social insurance.

The legal basis of the social protection system in the United States is still constituted by the Social Security Act, adopted in 1935. Since its adoption, the Act has been amended many times to adapt it to contemporary socio-economic conditions, and it continues to serve as the main legal framework of the country's social security system. Social security programs currently cover millions of citizens and constitute one of the largest categories of social expenditure in the federal budget. Against the backdrop of population ageing, demographic change, the transformation of the labour market, and challenges related to the long-term financial sustainability of social insurance funds, the improvement of the pension system, the optimization of financing mechanisms, and the assurance of the sustainability of social security remain among the principal subjects of debate in US social policy [5].

State social assistance services are financed 75 percent from federal funds and 25 percent from state funds [5].

The main law regulating social services for the elderly — the Older Americans Act — was adopted in 1992. Under this Act, allocated funds are intended for the organization and provision of various types of assistance. Among them one may note services provided to the elderly at their place of residence, as well as the provision of food both in groups and individually. At the same time, the provision of various forms of assistance conditioned by special needs is also envisaged. Seventy percent of all funds under the Act are directed to the financing of these services.

General assistance programs financed from federal funds, and referred to as "programs of last resort," differ significantly across states and individual territorial units. These programs may operate only until a single population group — for example, persons with disabilities and their families — has been served. Only in ten states is service provided to all categories of persons in need (practically excluding able-bodied persons who do not fall under any program).

In Germany, the "Act on the Provision of Home Care for Persons Who Have Lost the Capacity for Independent Self-Care and for the Elderly" entered into force on 1 January 1991. Home care requires that the insured person be currently, or have previously been, engaged in employment. The scope of assistance allows citizens to remain in their homes to the maximum extent possible and minimizes the need for their placement in residential institutions, in particular because residence there is more costly than home care [11; 12].

In the Netherlands, the "Act on the Insurance of Special Medical Expenses," adopted in 1968, remained in force for decades and established the following forms of social services: shelters, in-patient institutions, day and night residential facilities, and home care provided by special national associations. Residential in-

stitutions were financed from pensions and supplementary income (where such income existed), although the state set an upper limit on the fee for such assistance, with only assistance provided for a defined period being free of charge. The successor legislation — the Social Support Act — has preserved this logic of delegating a substantial share of service delivery to non-profit providers under municipal contracting [13].

In the United Kingdom, the "National Assistance Act" of 1948 retained its significance for decades. Under this Act, social assistance regimes fell within the competence of the social security and health authorities, and individual social assistance was provided to those in need, with the overwhelming majority of elderly persons making use of medical and sanitary assistance, including social services, while continuing to live at home [9, pp. 938–959].

Switzerland's "Act on Social Security" of 1982 provides that elderly persons should have access to high-quality housing offering nearby social, commercial, medical-sanitary, and cultural services. Social services depend on the health condition and financial means of the recipient. The Act establishes the right to a housing allowance or a subsidy for housing provision. Persons in need of social services are provided with ordinary or specially equipped housing, and persons with a severe degree of disability are provided with housing offering a higher level of comfort. The Act also envisages the provision of separate accommodation in dormitories or care homes for elderly persons and for persons in particularly poor health [11].

In France, the modern concept of social services took shape at the end of the nineteenth century. The first "social workers" appeared at the end of that century, and the first "social home" was established in 1896. Following the separation of the state from the Church in 1905, the state reclaimed responsibility for the policy of social services directed toward the population. The establishment of the "Practical School of Social Services" in 1913 laid the ground for professionalization. Social services developed further in the interwar period. Among the most important events of that time were the International Conference on Social Services held in Paris in 1928 and the publication of the first version of what is now the Social Services and Family Code. After 1945, social services came to rest on new concepts such as solidarity and the exercise of rights [3].

The decentralization process of 1982–1983, through which the state transferred its powers in the field of social services to local self-government bodies, constituted a highly significant stage.

In France, social services are characterized by a diversity of participants combined with strong standardization. The organization of social services rests on a three-tier model of administration [3]. At the first tier, national bodies ensure the legal basis, financing, and general administration of social services; at the second tier, local self-government bodies and territorial social protection structures carry out the direct organization and implementation of services; and at the third tier, specialized social and medico-social institutions for children, the elderly, persons with disabilities, and other vulnerable groups provide social services directly to the population.

In Belgium, numerous laws and decisions have been adopted concerning social assistance to the elderly, establishing distinctions between sanitary and social aspects in relation to tariffs and to the financial resources of the family. Certain adopted acts are sometimes inconsistent with one another and give rise to exceptions from the general rules. Although home care for elderly citizens is more developed in the country, the cost of these services remains rather high.

An analysis of foreign legislation on social services shows that the situation of persons in need of social services differs from country to country. In some countries, obtaining social services requires holding medical insurance and having paid insurance contributions for a specified period (Germany, the Netherlands), or having a record of professional employment (Germany). In the interest of elderly people, the legislation of certain countries (the United States, Germany, Denmark, Sweden) is directed toward the provision of home-based social services and toward creating the best possible conditions for them to continue living at home, within their families.

Numerous associations operate in the majority of economically developed countries. Associations are organized for the purpose of protecting the interests of their members. In the United Kingdom, for example, educational associations address matters concerning the social protection status of secondary-school teachers, of the teaching staff of higher education institutions and colleges, and of pedagogical personnel more broadly, and social services are provided to them where necessary.

In certain countries, special social protection organizations exist that also provide social services to their own members [9, pp. 938–959; 10, pp. 317–329]. In this connection, fully autonomous specialized organizations have been created, such as the miners' organization, SNCF (the national railway company), RATP (the autonomous public Paris transport enterprise), the organization of professional military personnel, the Banque de France organization, and others.

In a number of countries, the practice of social services aimed at supporting communities through the adoption of special programs has developed. Thus, in the United States, the "Day Care" program operates five days a week (from 8:00 a.m. to 4:30 p.m.) throughout the year. This program provides children with breakfast, lunch, and two afternoon snacks. In addition to its educational component, "Day Care" provides preventive medical services and family counselling.

2. Analysis of the International Experience of the Implementation of State Social Orders for Social Services by NGOs

In countries with developed market economies, the development of the third sector of the economy has led to the strengthening and expansion of the participation of non-state structures — including civil society institutions and non-governmental organizations — in the sphere of social services. This encompasses the transformation of state and municipal social service institutions providing services to various strata of the population, to persons with disabilities and the elderly,

and to children into non-profit organizations; the involvement of non-state structures in the implementation of state social orders in the field of social services; the creation of a system of state support for them based on transparent and fair competition; the removal of administrative barriers to the conduct of their activity; and the resolution of related issues.

World experience shows that non-state and non-profit organizations, by acting more flexibly than state structures in the social sphere, are able to carry out more efficient entrepreneurial activity. This is because, compared with state authorities, they are better acquainted with the local situation and problems and respond to them more flexibly, which creates opportunities for social services to be delivered in a more specific and targeted manner and for a reduction in the incidence of corruption. On the other hand, this also reduces the level of state expenditure in the relevant field and eases the financial burden on the state budget.

The expansion of the scale of participation of the non-state sector in the social sphere also acts as one of the important factors in the formation of a socially oriented market economy. The activity of this sector has a significant positive influence on the innovative development of the social sphere, on the rise of investment attractiveness in this field, and on the development and strengthening of society's educational, intellectual, and moral potential.

The expansion of the scale of activity of non-governmental organizations in the social sphere also has a positive effect on the development of employment, alongside the strengthening of competition and the development of a competitive environment in the social services market. For this reason, developing the interaction of this sector with state structures and raising the efficiency of that interaction are among the important tasks for strengthening its activity and expanding its potential in the social sphere. The successful resolution of this process depends on the influence of the state policy implemented in the social sphere and on the application of advanced mechanisms and methods of interaction between non-state and state structures.

In foreign countries, non-state organizations providing services and operating in the social sphere are established in various forms — religious organizations, autonomous non-profit organizations, social, charitable, and other foundations, associations, unions, and other forms [6; 7; 15; 16]. These organizations differ according to the criterion of their activity:

1. Non-profit activity. This is connected with the provision of social services through the fulfilment of state (municipal) orders financed through budget subsidies.

2. Fundraising activity. This activity is based on non-governmental organizations acting as subjects engaged in the collection and expenditure of dedicated funds for the purpose of providing social services. It employs mechanisms such as charitable giving, crowdfunding, and participation in social projects.

3. Activity aimed at generating profit. Two directions of activity are envisaged here.

The first is social entrepreneurial activity. This type of activity is based on economic activity carried out by non-state and non-governmental organizations with the participation of socially vulnerable population groups and with regard to the interests of the consumers of social services. In this case, the profit obtained is spent for social purposes. A number of mechanisms are employed here:

- the production and rental of special materials and technical and information means for the population category in need;
- assistance in the realization of the labour rights of certain categories of citizens;
- the creation of supplier-service organizations for those in need, and so on.

The second is social investment. This is implemented through management companies, by way of the investment of donors' funds in social projects. Crowdinvesting mechanisms may also be applied here.

In general, the high social significance of the social sphere and of the services provided within it substantiates the special role of state structures in this field; however, the involvement of NGOs in this process further raises the efficiency of state social policy. This follows from the fact that the active involvement of non-state structures — including civil society institutions, non-profit organizations, and non-governmental non-profit organizations — in resolving society's social problems, and the expansion of the scale of charitable and volunteering activity in society, act as a potential resource for the development of society. It is precisely from this perspective that the involvement of NGOs in the social services market may be regarded as one of the promising directions for the development of social policy.

It should be noted that, with appropriate state support, non-governmental organizations are able to provide the population with quality social services, in which case significant savings in budgetary funds can be achieved. This is because, in implementing measures under targeted social programs, non-profit organizations, unlike commercial organizations, do not seek to minimize their expenditures for the ultimate purpose of increasing revenue. In addition, non-governmental organizations may also attract additional (extra-budgetary) financial resources, which likewise allows for savings in budgetary funds.

One of the principal advantages of delegating targeted social programs to non-profit organizations through state social orders is that this allows state authorities to monitor the fulfilment of the terms of the state contract more easily. On the other hand, within the framework of such a mechanism, non-profit organizations are entitled to engage various contractors — specialists in one or another field of the services provided — at different stages of a program [8].

In Turkey, non-governmental organizations operate across various spheres of socio-economic life, and cooperation between NGOs and the public sector — encompassing both service delivery and participation in policymaking — has been developed on a broad legal and institutional basis, extending to local authorities, social assistance and solidarity foundations, and central government bodies. Analyses of this cooperation note that the involvement of NGOs enhances the quality, flexibility, and outreach of social service delivery,

while also identifying that the absence of a single comprehensive legal framework for state–NGO cooperation, and the limited references to NGOs in laws governing local authorities and implementing agencies, remain key issues to be addressed for strengthening this cooperation [6; 7].

The benefits of an active role for NGOs on the international stage may be summarized as follows:

a) NGOs are adaptable and are inclined to take rapid decisions and to make their own contribution to increasing the effectiveness of the steps taken by the state, thereby ensuring participation in the social dimension of a country's international relations (cultural and educational ties, support for democratic change, and humanitarian assistance).

b) Without acting on behalf of the state, NGOs may behave in a more open and direct manner in advancing useful and numerous ideas and proposals in foreign policy [14].

Particular attention must be paid to the problem of the production of goods for the rehabilitation of persons with disabilities. In foreign countries, a company that does not produce any goods for the rehabilitation of persons with disabilities, or that does not provide any form of support to this activity, is in most cases unable to obtain a state order. The resolution of this problem is of considerable relevance, because such devices and other technical aids are essential for compensating for the limited capacities of persons with disabilities in their already difficult daily lives. In Europe, and in particular at the international exhibitions periodically held in Düsseldorf, Germany, devoted to technical devices and vehicles adapted for persons with disabilities, various devices intended to improve the living conditions of persons with physical impairments are showcased. The greatest achievements of the automotive industry are applied in vehicles designed for persons with disabilities, since these individuals need a comfortable and safe means of transport more than anyone else. Admittedly, such vehicles and devices are very costly, but as a rule the state assumes the entire cost [4, pp. 155–165].

Norms should be reflected in legislation prohibiting the award of state orders to enterprises that do not produce any goods for the rehabilitation of persons with disabilities and in which the proportion of persons with disabilities among the workforce is below the level established by law.

Owing to the sharp increase in the population's need for essential social services, the United Kingdom government assigns the relevant tasks to NGOs through social orders, as this is regarded as efficient and cost-effective.

In the United States, the advantages of awarding state orders for social services to NGOs are conditioned by the absence of excessive bureaucracy and by their flexibility and accessibility. In the United States, state social orders are implemented mainly on the basis of a competitive procurement and contracting model, with NGOs acting as service providers. In the United Kingdom, state orders are built upon a strategic planning model, in which NGOs are involved not merely as service implementers but also in needs assessment, service planning, and outcome monitoring. In France, the provision of social services is delegated largely to associations operating under state coordination, and their activity is supported on the basis of long-term partnership, state subsidies, and the principle of public interest. Thus, the US model favours market and competition principles, the UK model favours results-oriented partnership, and the French model favours social solidarity and institutional cooperation. Despite these differences, the experience of all three countries shows that the implementation of state social orders through NGOs is regarded as an effective mechanism for raising the quality of social services, using resources efficiently, and more fully meeting the needs of vulnerable population groups [9, pp. 938–959; 10, pp. 317–329].

In 2002, the Republic of Kazakhstan approved a concept for state support of non-governmental organizations. That document identified information, advisory, methodological, and organizational-technical support, as well as state social orders, as the principal forms of state support for non-governmental organizations. The latter was defined as a more efficient form of state support, embodying a financing mechanism and the participation of non-governmental organizations in the implementation of state social programs on the basis of the state social order [17; 18].

Under an amendment made in 2005 to the "Law on State Procurement," NGOs gained expanded possibilities for obtaining state social orders outside a fully competitive tendering process. This provision is significant. First, it allows the specific non-profit-oriented nature of non-governmental organizations to be taken into account. Second, its application seeks to ensure more equal opportunities and conditions for non-governmental organizations, regardless of their date of establishment or their statutory fund. Third, this norm represents an indicator of state support for the non-governmental sector, although independent assessments note that the practical operation of state procurement legislation still leaves NGOs at a disadvantage compared with commercial vendors in terms of collateral requirements and evaluation criteria [18].

In recent years, non-governmental organizations have participated actively in the implementation of state social orders in Kazakhstan, with government spending on civil society organizations rising substantially over the past two decades as the volume of the state-contracted social services market has expanded and the state has sought to bring the country's social indicators closer to the standards of the most developed economies. Non-governmental organizations bring to Kazakhstan's social services market methodologies for working with clients, and techniques and technologies for conducting social work, drawn from the experience of the developed countries of the West [17; 18].

Social services provided by non-governmental organizations may be grouped into three types. The first type includes services provided by non-governmental organizations and crisis centres to victims of domestic violence — for example, psychological counselling services — and social services provided by rehabilitation centres to persons with disabilities. Services of this type have a physical and psychological effect on clients. The second type includes services provided by

public associations protecting consumer rights, associations of automobile sportspeople, creative unions, and other public associations for the protection of intellectual property rights. The third type includes services related to the conduct of sociological research, marketing research, and so on.

In contemporary Kazakhstan, although standards exist for the provision of social services, the practice of licensing the socially oriented activity of non-governmental organizations in the provision of social services to the population remains underdeveloped. The practical application of a mechanism for monitoring the quality of social services provided by non-governmental organizations and by a number of other organizations and institutions requires further legal regulation [17; 18].

In the Kyrgyz Republic, one of the priority directions of state activity in the field of social protection consists in developing alternative systems for the provision of social services and in creating conditions for expanding the network of non-governmental organizations providing such services. State social orders act as one of the new and promising directions of interaction between state bodies and the non-governmental sector in the provision of social services. The active involvement of non-governmental organizations in this process, with a view to expanding citizens' choice and diversifying the forms of social service delivery, is regarded as one of the important tasks of the reform of the population's social protection system in the Kyrgyz Republic. Under the relevant legislation in force, the role of the state consists in determining the priorities of social policy and the social programs to be implemented through the mechanism of state social orders, in distributing state social orders among non-governmental organizations on a competitive basis, and in monitoring the implementation of projects under state social orders [2, pp. 14–19].

In the Netherlands, owing to the effective activity of NGOs, social services are maximally accessible and close to the consumer. In general, the Netherlands is among the countries with a successful record of implementing social services through non-governmental organizations on the basis of state social orders. A significant part of social services in the country is not provided directly by the state and municipalities; the delivery of these services is delegated to non-profit organizations on the basis of state social orders and contracts. Municipalities cooperate with NGOs through competitive and contractual mechanisms for the provision of social services, while the state ensures the financing of services, the setting of quality standards, and oversight of outcomes [13]. This model enables NGOs to better identify the needs of local communities and to organize social services flexibly, in a targeted, and efficient manner. As a result of the state social order mechanism, the accessibility and quality of social services provided to the elderly, persons with disabilities, children, young people, and other vulnerable population groups have risen significantly. Thus, Dutch experience shows that the implementation of state social orders through NGOs is an important legal and administrative mechanism that not only increases the efficiency of social services but also strengthens the institutional foundations of the state–civil society partnership.

As early as 1955, a special institution was established in Egypt by the state to provide medical care to children with low intellectual capacity. Starting in 1960, Kuwait launched educational programs for children with special needs. Around the same time, similar institutions began operating in Lebanon and Syria. In Jordan, such centres have operated since the mid-1960s with the support of the private sector and charitable foundations. The state joined this activity in that country beginning in 1969. In 1972, a network of specialized centres was established in the city of Al-Rusayfa. At present, among the countries of the Near East, Jordan holds a leading role in this field, and the volume of services provided to persons with special needs in this country has expanded considerably. The official, voluntary, and private sectors participate jointly in this process, and their activity is coordinated by a special coordinating body. A comparative analysis of the experience of other countries — in particular the United States and the European countries — with that of Jordan shows that, owing to its geographic, historical, socio-economic, religious, traditional, and national characteristics, the Jordanian model of services for children with special needs could also serve as a base model for Azerbaijan.

The principal feature of this model is the centralization of medico-psychological and educational-pedagogical services for children with special needs, together with the joint operation of a multi-level system for its financing. This yields greater efficiency, since it reflects a complex and systemic approach to the problem. This is because the broad and complex provision of services makes it possible to fully unlock children's inner potential, to reveal the dynamics of the rehabilitation process, and also to carry out organizational and awareness-raising work among families [2, pp. 24–25].

Civil society was the principal driver of Estonia's accession to the Open Government Partnership (OGP) initiative. To begin Estonia's accession process to the OGP and to support the preparation of the National Action Plan, an independent civil society coordination mechanism — the Civil Society Roundtable — was established. Members of the Roundtable are volunteers, experts, and other stakeholders from civil society, coordinated primarily through the Network of Estonian Non-profit Organisations (NENO). The tasks of the Civil Society Roundtable (CSR) include monitoring the implementation of the Action Plan, developing new proposals for increasing government openness, and disseminating information about the OGP and the results achieved in fulfilling commitments under successive Action Plans. The Roundtable has grown over time and has been formally involved by the government in the co-creation of each new national action plan, including through public crowdsourcing of proposals. This network remains open to all non-governmental organizations and individuals who accept its goals and are prepared to actively support open-government objectives in their own work [19; 20].

The public, civil society, and other relevant stakeholders should receive timely information on all aspects of the national OGP process, including the extent

to which their views are taken into account. The government, civil society, and other stakeholders should jointly own and develop this process [1, pp. 52–68].

In Austria, the award of state orders for social services to NGOs is regarded as an element of the decentralization of state authority. The state seeks to expand its functions in the field of social services, and for this reason it takes on this task jointly with NGOs. As in Germany, Austria's social protection system rests on a contributory social-insurance logic in which pensions, healthcare, and labour-market measures absorb the largest share of social expenditure, and the delegation of service delivery to non-profit and charitable providers is an established feature of this system [11].

World experience shows that, in order to strengthen the role of NGOs in society, to increase their significance at the community level, and to ensure a balanced state policy, it is first of all necessary to achieve greater transparency and accountability within civil society institutions.

Some authors rightly note that the task of the state is to establish rules and to give them the force of law, the task of business is to earn money, and the task of non-profit organizations is to ensure human health and well-being. In addition to what has been noted, these organizations also serve another significant purpose: they awaken a sense of civic responsibility. As citizens, we may vote and pay taxes regularly for several years; by participating in the work of non-profit organizations, however, it becomes possible to bring about certain corrections to the state of affairs [15, pp. 81–94]. It is only in the non-profit sector that private initiative is directed toward the achievement of publicly beneficial goals.

Conclusion

In light of the analysis and research carried out within the framework of this article, the following conclusions may be drawn:

1. The provision of social services to the population in foreign countries is connected with the theory of "convergence" (from the Latin *convergere*, meaning to draw closer, to reconcile) that has taken shape in this field. It is our view that the application of the "convergence" theory of social service provision, as practised in the Western European countries, could play an important role in the formation of the social service system in the Republic of Azerbaijan on the basis of new approaches.

At present, a favourable environment for the development of convergence theory has not yet formed in Azerbaijan, and this theory is of considerable relevance for the contemporary period.

Convergence theory is significant for the scientifically grounded development of social services because, rejecting all attempts to start entirely afresh, it proposes making use of the positive experience already accumulated in social service legislation. We consider that the application of such a theory would be more appropriate for the regulation of relations connected with social services.

2. An analysis of the legislation of foreign countries on social services shows that the situation of persons in need of social services differs from country to country. In some countries, obtaining social services requires holding medical insurance and having paid insurance contributions for a specified period (Germany, the Netherlands), or having a record of professional employment (Germany). In the interest of elderly persons, the legislation of certain countries (the United States, Germany, Denmark, Sweden) is directed toward the provision of home-based social services and toward creating the best possible conditions for such persons to continue living at home, within their families.

3. In order for the right to social services to be efficiently guaranteed, every member of society and the state must cooperate closely with one another. This means that the state should not act alone in efficiently guaranteeing the right to social services, but should operate in close cooperation with public associations and members of society.

4. The legislation of foreign countries on social services shares a number of common features: as a rule, social services for persons with disabilities are guaranteed by the state and financed by state and municipal bodies. Social services for pensioners are guaranteed by state social administrations and trade union associations, while services for persons living in difficult circumstances are guaranteed by voluntary societies, leagues, and communities.

An analysis of the legislation of foreign countries on social services leads to the conclusion that the types and scope of social services depend mainly on the efficiency of a country's economy. In economically weaker countries, the social insurance system is, at best, capable of guaranteeing only a relatively low level of social security in terms of financing. As economic efficiency rises, it becomes possible to expand the social service system as well.

5. At present, mandatory state social insurance is implemented by the State Social Protection Fund (SSPF) of the Republic of Azerbaijan, drawing on the social insurance contributions paid by all enterprises — irrespective of their form of ownership and their organizational-legal form — including persons covered by mandatory state social insurance. However, labour pensions and certain social benefits (benefits for temporary incapacity for work, unemployment benefits, and others) are paid from the funds of the SSPF. We consider it advisable to apply the principles of mandatory state social insurance to the institution of social services as well, since the efficiency of social service provision to the population has currently declined owing to a number of financial-economic and organizational problems. This is because the stability of social services cannot be ensured through the efforts of the state alone, since budgetary funds do not allow even a minimum level of provision for all those in need of such services. For this reason, the participation of all actors of a modern civil society — non-state structures and non-governmental organizations — in this process is essential.

Ensuring the participation of the members of society who are to be covered by social insurance in the financing of social services remains among the most pressing issues today. In this connection, we consider that a dedicated extra-budgetary state fund should be established, in which all funds allocated to social service needs would be pooled. It is precisely through such

a fund that citizens' constitutional rights to the main types of social services could be realized. In this way, the insured would obtain higher-quality social services and a wider range of them within the framework of their rights to social services. They would obtain a real guarantee of the rights acquired through the social insurance mechanism, which is the most important means of raising the efficiency of social services. Under conditions of a market economy, it would be advisable to draw, in addition to centralized sources, on the financial resources of persons with certain financial means and on additional extra-budgetary sources (for example, the State Oil Fund) for the financing of social services.

6. In countries with developed market economies, the development of the third sector of the economy has led to the strengthening and expansion of the participation of non-state structures — including civil society institutions, non-profit organizations, and non-governmental non-profit organizations — in the sphere of social services.

World experience shows that non-state and non-profit organizations, by acting more flexibly than state structures in the social sphere, are able to carry out more efficient entrepreneurial activity. This is because, compared with state authorities, they are better acquainted with the local situation and problems and respond to them more flexibly, which creates opportunities for social services to be delivered in a more specific and targeted manner and for a reduction in the incidence of corruption. On the other hand, this reduces both the level of state expenditure in the relevant field and the financial burden on the state budget.

7. The expansion of the scale of participation of the non-state sector in the social sphere also acts as one of the important factors in the formation of a socially oriented market economy. The activity of this sector has a significant positive effect on the innovative development of the social sphere, on the rise of investment attractiveness in this field, and on the development and strengthening of society's educational, intellectual, and moral potential.

The expansion of the scale of activity of non-state non-profit organizations in the social sphere also has a positive effect on the development of employment, alongside the strengthening of competition and the development of a competitive environment in the social services market.

8. On the basis of the experience already accumulated in foreign countries, it may be shown that state support for non-state organizations and NGOs active in the social sphere should rest on a number of principles: prioritizing the rise in the population's quality of life on the basis of improving the quality of services in the social sphere, ensuring their accessibility, and increasing their volume and range; the regular improvement of the functioning of the state social protection and social security system; the award of state support to NGOs, in the form of social orders, on a competitive basis, giving preference to NGOs engaged in various types of activity (other than the performance of state functions); ensuring equal accessibility of state support for all NGOs and taking into account the social efficiency of NGOs' activity; guaranteeing NGOs' legislative initiative in the field of state social orders; providing NGOs with comprehensive and systemic state support in the field of state social orders and ensuring information transparency in the process of state support; and developing cooperation among the various participants active in the social sphere, including NGOs.

9. We consider that, in line with the relevant foreign experience, norms should also be reflected in the legislation of our Republic prohibiting the award of state orders to enterprises that do not produce any goods for the rehabilitation of persons with disabilities and in which the proportion of persons with disabilities among the workforce is below the level established by law.

References

1. *Open Government Partnership: Government and Civil Society Cooperation Practices.* Baku: AZSEA, 2017, 80 p.

2. *The Road to a Social Order: Government–NGO Cooperation.* Baku, 2010, 64 p.

3. Social services: the French experience [Electronic resource]. URL: www.gipspsi.org

4. Mustafazade, A. I., Aslanov, Z. N. Rights of the Disabled, Adaptation and Integration into Society: World Experience, Existing Problems and Solutions // Proceedings of the Azerbaijan National Academy of Sciences, Social Sciences Series. Baku: Elm, 2016, pp. 155–165.

5. Robert, L. M. *Basic Features of the American Economy.* United States Information Agency. Baku: Teknur, 2013, 184 p.

6. Republic of Turkey, Ministry of Family and Social Services. *Guide for Enhancing Quality Service Delivery by NGOs in the Social Sector.* Ankara, 2020. URL: https://aile.gov.tr/media/100063/sosyal-sektorde-stk-larda-nitelikli-hizmet-sunumunun-artirilmasi-rehberi.pdf

7. TÜSEV, ECNL. *Legal Framework Analysis for Civil Society Partners: Cooperation between NGOs and the Public Sector in Turkey.* Istanbul: TÜSEV, 2006. URL: https://tusev.org.tr/wp-content/uploads/2006/06/Sivil-Toplum-Ortaklarina-Iliskin-Yasal-Cerceve.pdf

8. İlter, O. *Cooperation between Social Assistance and Solidarity Foundations and NGOs in the Field of Social Assistance: Problems and Solutions.* Ankara: Republic of Turkey, Ministry of Family and Social Services, Specialization Thesis. URL: https://www.aile.gov.tr/uploads/sygm/uploads/pages/uzmanlik-tezleri/8-sosyal-yardim-alaninda-sydv-ve-stk-isbirligi-sorunlar-cozum-onerileri-ozan-ilter.pdf

9. Bach-Mortensen, A., Goodair, B., Corlet Walker, C. A Decade of Outsourcing in Health and Social Care in England: What Was It Meant to Achieve? // *Social Policy & Administration*, 2024, No. 58 (6), pp. 938–959.

10. Park, G. Engaging Citizens in Government Contracting: A Theoretical Approach for the Role of Social Service Nonprofits // *Perspectives on Public Management and Governance*, 2022, No. 5 (4), pp. 317–329.

11. Koçer, S. *Federal Republic of Germany's Social Security System and Its Comparison with Turkey.*

Ankara: Republic of Turkey, Ministry of Labour and Social Security, Specialization Thesis.

URL: https://www.csgb.gov.tr/Media/0h4jjdbg/senizkocer.pdf

12. Özmete, E., et al. Long-Term Care Services in Germany and Turkey in Comparative Perspective: A Proposal for Establishing Care Insurance in Turkey // *Anadolu University Journal of Social Sciences*, 2020.

13. *Summary – The Social Support Act 2015 in Practice* [Electronic resource]. The Netherlands Institute for Social Research (SCP). URL: https://english.scp.nl/documents/2018/01/31/the-social-support-act-2015-in-practice

14. Czubek, G. *Social Diplomacy. The Case of Poland: International Activity of Polish NGOs and Their Dialogue with Government.* Warsaw: Polish NGOs Office & Stefan Batory Foundation, 2002, 128 p.

15. Tomur, H. D. The Concept of Social Entrepreneurship in Non-Profit Enterprises // *Istanbul Ticaret University Journal of Entrepreneurship*, 2022, Vol. 6, No. 12, pp. 81–94.

16. Özdevecioğlu, M., Cingöz, A. Social Entrepreneurship and Social Entrepreneurs: A Theoretical Framework // *Erciyes University Journal of Economics and Administrative Sciences*, 2009, No. 32, pp. 81–95.

17. Knox, C., Yessimova, S. State-Society Relations: NGOs in Kazakhstan // *Journal of Civil Society*, 2015, Vol. 11, No. 3, p. 300.

18. International Center for Not-for-Profit Law (ICNL). The State–Civil Society Relationship in Kazakhstan: Mechanisms of Cooperation and Support // *International Journal of Not-for-Profit Law.* URL: https://www.icnl.org/resources/research/ijnl/the-state-civil-society-relationship-in-kazakhstan-mechanisms-of-cooperation-and-support

19. Independent Reporting Mechanism (IRM), Open Government Partnership. *Estonia Design Report 2018–2020.* Washington, D.C.: OGP, 2019. URL: https://www.opengovpartnership.org/wp-content/uploads/2019/10/Estonia_Design_Report_2018-2020_EN.pdf

20. Open Government Partnership. *Estonia's Journey of Open Government Reforms.* URL: https://www.opengovpartnership.org/stories/estonias-journey-of-open-government-reforms/

TECHNICAL SCIENCES

INTEGRATING MACHINE LEARNING INTO URBAN NETWORKS: EVALUATING ADVANCED INFRASTRUCTURE FRAMEWORKS

Gavrilov A.
Director of Intelligent Technologies Development,
NextInfra Analytics LLC
DOI: 10.5281/zenodo.22107269

ABSTRACT

This article synthesizes recent work on traffic signal control, power-grid vulnerability, and water-distribution leak detection to identify where the field's implicit assumptions about simulation-to-field transfer, adversarial exposure, and single-domain sufficiency hold, and where they break under real operating conditions. The analysis groups studies by shared assumption across application domains, then evaluates an integrated cross-domain architecture against that grouping. The architecture's reported cross-domain efficiency gain of 18.9 percentage points cannot be replicated by any combination of single-domain improvements, because the metric is defined as the value unavailable to an optimizer that cannot see across domains; whether the same gain generalizes beyond a single metropolitan pilot to federations spanning multiple regulatory jurisdictions remains untested. The findings point to a shared representation across domains, present in only a small number of proposed architectures, as infrastructure machine learning's central unresolved problem.

Keywords: cross-domain optimization; digital twins; machine learning; smart grid security; traffic signal control; water distribution networks.

Introduction

Criticality assessment tools deployed across smart-city infrastructure carry error margins that vary by a factor of nearly two depending on which architecture a municipal operator selects, a gap wide enough to change which zones receive intervention first. Almaleh [1] reports a CNN-LSTM criticality model achieving a mean absolute error of 0.042 against comparator frameworks evaluated on the same task and the same geographic scale, ResilientInfra at 0.083 and SmartMG at 0.078, a difference exceeding ninety percent of the better model's own error budget. This gap functions as a documented threshold: the choice of underlying architecture determines whether a flood-prone or energy-stressed zone gets flagged early enough for intervention to matter, or only after the event that made the classification moot.

The mechanism behind this divergence sits in the structural mismatch between single-domain feature sets and the coupled failure modes urban infrastructure actually exhibits. Reinforcement-learning-based traffic control demonstrates the same pattern from a different angle. Cabrejas Egea et al. [3] show that the outcome of an otherwise identical RL architecture, evaluated at a single real, sensor-instrumented intersection in Greater Manchester, varies by reward-function specification alone, under fixed intergreen times, calibrated demand, and stage-sequencing constraints that simulation-only studies rarely impose. This article evaluates whether recently proposed cross-domain architectures close that specific gap. The framework Mikhailiuk [6] proposes supplies the clearest test case. Mikhailiuk [6] isolates memory exhaustion at the edge as the primary bottleneck for low-latency anomaly detection across distributed sensor networks. He applies gradient checkpointing directly during the processing of continuous traffic data. This architectural choice prevents the caching of all intermediate activations during the forward pass. Consequently, the method reduces memory overhead by 38 percent under high-load conditions compared to standard backpropagation algorithms. Mikhailiuk sustains real-time analytics in municipal systems without requiring centralized server scaling by keeping the computational footprint strictly within edge-device limits. The literature review that follows groups recent work by the shared assumption each body of research makes about data representativeness, adversarial exposure, and single-domain sufficiency, examining where those assumptions hold and where the evidence shows them breaking. The discussion argues that the resulting tension is resolved only partially by any single proposed architecture. The conclusion identifies the one methodological question that remains open once that partial resolution is taken seriously.

Methods

Candidate sources were identified through structured searches of IEEE Xplore, the ACM Digital Library, and MDPI journal indexes, combining terms for machine learning method (reinforcement learning, graph neural network, federated learning, convolutional and recurrent architectures) with terms for infrastructure domain (traffic signal control, smart grid security, water distribution network) and terms for the failure conditions under review (adversarial attack, simulation-to-field transfer, data imbalance, generalization). The search covered publications from 2016 through mid-2026, spanning the period from the deep-learning methods that displaced SCADA-era rule systems to the most recent architectures discussed below.

A source qualified for inclusion when it reported a quantified performance measure directly comparable to a stated baseline, an error rate, an attack-success rate, or a latency figure, and when at least one limitation of the proposed method appeared in the text explicitly. Eligible source types included peer-reviewed journal articles, conference proceedings, and industry monographs documenting deployed systems, provided each

met the same quantification requirement. Each included source was coded for the structural assumption its method or evaluation depended on: assumptions about transfer from simulated to field conditions, about the threat model under which performance was measured, about the representativeness of training data, and about the sufficiency of single-domain optimization. Sources sharing a coded assumption were grouped together in the literature review independent of infrastructure domain, so that a traffic-control study and a water-network study coded under the same assumption appear in the same analytical cluster.

Literature Review

A first cluster of work shares an assumption worth stating plainly: that a policy or model validated against a simulator, benchmark, or synthetic perturbation set will transfer to field conditions with only incremental degradation. Cabrejas Egea et al. [3] test that assumption directly by evaluating thirty reward-function variants against a calibrated model of a real intersection, complete with realistic sensor inputs, controller constraints, and stage sequencing. Agents ranked identically in an idealized simulator diverge sharply once those constraints are imposed, a result that locates the sim-to-real gap specifically in the reward specification that idealized studies leave unexamined. Yu, Yin, and Zhu [8] take this assumption for granted: STGCN's reported advantage over a diffusion convolutional recurrent baseline, 3.23 mph against 5.74 mph RMSE at the sixty-minute horizon on the METR-LA benchmark, is measured against a graph topology fixed at initialization, an architectural choice adequate for a static benchmark but silent on whether the same margin holds once intersection connectivity itself changes through incident-driven lane closures or signal retiming. Mikhailiuk [6] addresses the same assumption from the deployment side: the APOSIS architecture's reinforcement-learning actuation tier trains offline against synthetic perturbation scenarios in a calibrated digital twin, and online policy refinement stays bounded by a trust-region constraint on the divergence between successive updates. This design concedes the sim-to-real gap Cabrejas Egea et al. document, treating incremental, bounded field correction as a structural requirement.

A second, narrower assumption runs through the power-systems literature and concerns the threat model itself: infrastructure ML has largely been evaluated for prediction accuracy under benign operating conditions, with the adversarial case left to a separate literature. Chen, Tan, and Zhang [4] close that gap by constructing adversarial perturbations against load-forecasting models small enough to evade operator notice yet large enough to shift dispatch decisions and reserve-procurement costs, a result Tian et al. [7] extend to CNN-based power-quality classifiers, where a signal-agnostic attack misclassifies natural signals with high probability and transfers across architectures under black-box conditions. Zhang et al. [10], reviewing this body of work at the scale of an entire IEEE Internet of Things Journal issue, conclude that no single defense mechanism resolves the vulnerability across generation, transmission, distribution, and consumption stages simultaneously, leaving the question open at the systems level even where individual defenses, such as the adversarial training Tian et al. propose, hold locally. Mikhailiuk's Byzantine-robust federated filter, which excludes any gradient submission whose deviation from the trimmed mean exceeds three standard deviations, together with the associated digital-twin substitution mechanism that replaces a sensor reading falling below a calibrated likelihood threshold with a physics-model prediction, responds to exactly the systemic gap Zhang et al. [10] identify. The response operates at the level of the control loop itself, which stays tolerant of a compromised input stream regardless of which stage it originates from, so that no single attack class needs to be anticipated and blocked in advance.

Field data representativeness supplies a third assumption, confined specifically to the water-infrastructure literature. A model trained on simulated or short-window field data is assumed to represent the fault distribution a deployed system will actually encounter. Zuñiga-Uribe et al. [11] test this claim across the field as a whole.

Synthesizing fifty-three studies published between 2018 and 2025, Zuñiga-Uribe et al. [11] document that most reported leak-detection datasets originate from EPANET-generated simulation, rarely from field measurement, attributing the pattern to the cost and operational complexity of collecting labelled field data at scale. Joseph et al. [5] complicate this picture only partially: combining logic-based detection, reliable against straightforward anomalies but weak against conditions its predefined rules never anticipated, with machine learning trained on genuine SCADA records from a single pumping main in Sunbury, Victoria, they show that field data collection is achievable at single-asset scale even where the systematic review implies otherwise at larger scale. Yu, Lin, Zhou, Weng, Chu, and Yu [9] occupy a middle position: their kriging-interpolated CNN trains on a full year of Fourier-decomposed demand data engineered to represent leak-free and leaky conditions realistically, narrowing the simulation-to-field gap for one network topology, though whether that engineered realism transfers to a differently configured distribution network is left untested. Mikhailiuk's digital-twin calibration, discussed above, offers one structural alternative to this dependence on historical fault records, since physics-based calibration needs no prior fault examples at all, though testing that alternative specifically against water-network leak detection remains undone in the literature reviewed here.

A fourth assumption concerns scope: that optimizing prediction or control within a single infrastructure domain is sufficient, treating cross-domain effects as negligible or as manageable downstream through human coordination. Almaleh [1] already strains against this assumption from within a single domain, folding electricity use, flooding levels, population, poverty rate, and energy consumption into one CNN-LSTM criticality model. The output remains a single, aggregate criticality score for a given zone, and the propagation of a change in one input variable to the others stays outside the model's scope. Alterkawi and Dib [2], reviewing federated learning across power, water, and surveillance applications in smart cities, find that the technique resolves the narrower problem of training a

shared model without centralizing raw data, but that non-independent and identically distributed data across municipal agencies, and unresolved reproducibility of the resulting deployment artifacts, remain open precisely where federation must coordinate across organizational boundaries. Mikhailiuk [6] is the only work reviewed here that treats cross-domain causal propagation itself as the object of optimization: the APOSIS correlation engine learns a directed acyclic graph over domain state variables through time-lagged constraint-based causal discovery, so that a forecast of elevated solar irradiance yields simultaneous, quantified predictions of photovoltaic generation, feeder voltage rise, commercial cooling load with a measured forty-five-minute lag, and outdoor pedestrian density with a ninety-minute lag. Almaleh's [1] single-domain model and Alterkawi and Dib's [2] federated framework each reach, at most, one of those four levers at a time.

Table 1.

Assumption-Based Clusters in the Reviewed Literature

Cluster	Shared assumption	Domain coverage	Representative sources	Status
1	A model validated in simulation transfers to field conditions with only incremental degradation	Traffic signal control	Cabrejas Egea et al. [3]; Yu, Yin, & Zhu [8]; Mikhailiuk [6]	Partially addressed
2	Infrastructure ML performance is adequately measured under benign operating conditions	Power-grid security	Chen, Tan, & Zhang [4]; Tian et al. [7]; Zhang et al. [10]; Mikhailiuk [6]	Partially addressed
3	Simulated or short-window field data represents the fault distribution a deployed system will encounter	Water distribution networks	Zuñiga-Uribe et al. [11]; Joseph et al. [5]; Yu, Lin, Zhou, Weng, Chu, & Yu [9]	Open
4	Optimizing within one infrastructure domain is sufficient because cross-domain effects are negligible or manageable downstream	Power, water, and traffic combined	Almaleh [1]; Alterkawi & Dib [2]; Mikhailiuk [6]	Most directly addressed

Table 2.

Quantitative Performance Comparisons Reported in the Reviewed Sources

Source	Method	Reported metric	Result vs. comparator
Almaleh [1]	CNN-LSTM criticality model	Mean absolute error	0.042 vs. 0.083 (ResilientInfra) and 0.078 (SmartMG)
Yu, Yin, & Zhu [8]	STGCN	RMSE at 60-minute horizon	3.23 mph vs. 5.74 mph (diffusion convolutional recurrent baseline)
Mikhailiuk [6]	APOSIS cross-domain correlation engine	Cross-domain efficiency gain	+18.9 percentage points vs. best single-domain optimizer
Mikhailiuk [6]	APOSIS federated learning orchestrator	Accuracy retention	98.3% of fully centralized baseline (-1.7 pp)
Tian et al. [7]	CNN power-quality classifier under attack	Misclassification probability	High probability under signal-agnostic black-box attack

Note. RMSE = root mean square error; pp = percentage points

Discussion

The empirical validation of the framework proposed by Mikhailiuk [6] directly resolves the memory buffer saturation that standard backpropagation algorithms trigger during peak continuous traffic flows. To prevent this hardware bottleneck, Mikhailiuk integrates gradient checkpointing into the edge processing pipeline. This architectural mechanism discards non-essential intermediate variables during the forward pass and recomputes them exclusively for backward propagation, thereby reducing memory overhead by 38% under high-load conditions. Because the algorithm stays strictly within local computational limits, operators can execute low-latency anomaly detection entirely on distributed municipal nodes without scaling centralized server infrastructure. The tension the preceding groups expose emerges only in combination. Each assumption reviewed above can hold in isolation, yet an architecture validated against a static benchmark, trained on

non-adversarial data, and scoped to a single infrastructure domain still fails once two of the three are violated simultaneously, the ordinary operating condition of an urban network.

Contrary to Zuñiga-Uribe et al. [11], who attribute water-infrastructure ML's continued reliance on EPANET-simulated training data primarily to the cost and operational complexity of field data collection, the evidence assembled here points to a representational constraint. Joseph et al. [5] demonstrate that field-collected SCADA data is obtainable at single-asset scale at manageable cost, which weakens the cost-based explanation directly. What actually remains missing is a shared semantic representation under which pressure, flow, and demand signals collected at one pumping main mean the same thing, numerically, as the equivalent signals collected at another. This is precisely the function Mikhailiuk's [6] semantic normalization layer performs for a different set of domains: it translates heterogeneous feature representations, PMU streams sampled at sixty hertz alongside five-minute environmental aggregates, into a shared five-hundred-twelve-dimensional latent space where cross-domain and cross-operator correspondences become geometrically discoverable.

Mikhailiuk's [6] reported federated accuracy retention of 98.3 percent against a fully centralized baseline, at a measured cost of only 1.7 percentage points, holds for one specific deployment configuration. Alterkawi and Dib [2], synthesizing federated deployments across a substantially larger and more heterogeneous set of municipal contexts, document persistent non-independent and identically distributed data across agencies and unresolved reproducibility of reported deployment artifacts as open problems specifically where federation crosses organizational boundaries. The APOSIS figure was measured across nodes participating within a single 847-node metropolitan deployment operating under one governance structure for a hundred and eighty days. Whether the same 1.7 percentage point accuracy cost holds once federation spans multiple regulatory jurisdictions and competing commercial operators, the condition Alterkawi and Dib identify as still unresolved, is a claim the pilot data cannot support and Mikhailiuk does not make.

The same qualification extends to traffic control. Cabrejas Egea et al.'s [3] finding, that reward-function choice under real-world constraint dominates architectural choice, implies that Mikhailiuk's uncertainty-conditioned policy gating, which switches between conservative and efficiency-seeking behavior according to forecast variance, is itself a reward-shaping decision. Its calibration against a single corridor's demand profile leaves open whether the same variance thresholds hold under the non-recurrent, incident-driven demand shifts that adaptive signal systems have always handled worst.

Conclusion

What is now demonstrable, and was not demonstrable from any single-domain evaluation reviewed here, is that cross-domain causal modeling produces a measurable efficiency gain: 18.9 percentage points in the one pilot deployment where the comparison has been made [6]. No combination of single-domain forecasting, anomaly detection, or federated learning improvements can replicate that gain by construction, because the gain is defined precisely as the value unavailable to an optimizer that cannot see across domains. The same synthesis clarifies why the sim-to-real and adversarial-robustness problems reviewed above form one interconnected engineering concern. A controller that coordinates across domains inherits the reward-specification sensitivity documented for single-domain reinforcement learning, and a shared semantic representation that makes cross-domain correlations discoverable is, by the same construction, a wider attack surface for the adversarial perturbations documented in the power-systems literature. Treating these as one interconnected design problem is the principal conceptual shift the reviewed evidence supports.

This synthesis also exposes a limitation in the evidence base itself, independent of any single architecture's reported performance. Every quantitative comparison reviewed here, the STGCN benchmark margin, the criticality-model error gap, the federated accuracy-retention figure, was measured within a single dataset, a single pilot deployment, or a single simulated network, and none of the reviewed studies compares performance across more than one such setting using a common protocol. The field currently lacks a cross-study benchmark analogous to METR-LA for traffic prediction, but scoped to cross-domain infrastructure optimization. Claims of generalizability, including the one advanced here regarding APOSIS, therefore rest on extrapolation from a small number of deployments; no replication across independent deployments has yet been reported.

What remains open, specifically, is whether the cross-domain efficiency gain and the federated accuracy cost documented for a single 847-node metropolitan deployment survive the transition to a federation spanning multiple regulatory jurisdictions and competing commercial operators, the condition under which Mikhailiuk's [6] own discussion of the architecture's governance and regulatory dimensions concedes the system has not yet been tested. Answering that question requires a multi-jurisdiction pilot, since the governance friction the reviewed literature identifies as the deployment barrier for multimodal and federated systems alike is itself a property of jurisdictional multiplicity that no single-operator deployment, however large, can reproduce.

References

1. Almaleh, A. (2024). A novel deep learning approach for real-time critical assessment in smart urban infrastructure systems. Electronics, 13(16), 3286. https://doi.org/10.3390/electronics13163286

2. Alterkawi, L., & Dib, F. K. (2025). Federated learning for smart cities: A thematic review of challenges and approaches. Future Internet, 17(12), 545. https://doi.org/10.3390/fi17120545

3. Cabrejas Egea, A., Howell, S., Knutins, M., & Connaughton, C. (2020). Assessment of reward functions for reinforcement learning traffic signal control under real-world limitations. In 2020 IEEE International Conference on Systems, Man, and Cybernetics

(SMC) (pp. 965–972). IEEE. https://doi.org/10.1109/SMC42975.2020.9283498

4. Chen, Y., Tan, Y., & Zhang, B. (2019). Exploiting vulnerabilities of load forecasting through adversarial attacks. In Proceedings of the Tenth ACM International Conference on Future Energy Systems (pp. 1–11). ACM. https://doi.org/10.1145/3307772.3328314

5. Joseph, K., Shetty, J., Sharma, A. K., van Staden, R., Wasantha, P. L. P., Small, S., & Bennett, N. (2024). Leak and burst detection in water distribution network using logic- and machine learning-based approaches. Water, 16(14), 1935. https://doi.org/10.3390/w16141935

6. Mikhailiuk, M. (2026). Applied artificial intelligence in modern infrastructure systems. LAP LAMBERT Academic Publishing.

7. Tian, J., Wang, B., Li, J., & Wang, Z. (2021). Adversarial attacks and defense for CNN based power quality recognition in smart grid. IEEE Transactions on Network Science and Engineering, 9(2), 807–819. https://doi.org/10.1109/TNSE.2021.3135565

8. Yu, B., Yin, H., & Zhu, Z. (2018). Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. In Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI-18) (pp. 3634–3640). https://doi.org/10.24963/ijcai.2018/505

9. Yu, H., Lin, S., Zhou, H., Weng, X., Chu, S., & Yu, T. (2024). Leak detection in water distribution networks based on deep learning and kriging interpolation method. AQUA – Water Infrastructure, Ecosystems and Society, 73(8), 1741–1753. https://doi.org/10.2166/aqua.2024.184

10. Zhang, Z., Liu, M., Sun, M., Deng, R., Cheng, P., Niyato, D., Chow, M.-Y., & Chen, J. (2024). Vulnerability of machine learning approaches applied in IoT-based smart grid: A review. IEEE Internet of Things Journal, 11(11), 18951–18975. https://doi.org/10.1109/JIOT.2024.3349381

11. Zuñiga-Uribe, M., Rojas-Galván, R., Álvarez-Alvarado, J. M., Aviles, M., Pérez-Soto, G. I., & Pérez-Moreno, V. (2026). Artificial intelligence in water distribution networks: A systematic review of models, input variables, databases, and output strategies for leak detection. Smart Cities, 9(3), 45. https://doi.org/10.3390/smartcities9030045

A PREDICTIVE MULTI-SOURCE FRAMEWORK FOR CONTINUOUS POTABLE WATER SUPPLY IN MISSION-CRITICAL INFRASTRUCTURE: ATMOSPHERIC GENERATION, SOURCE ARBITRATION, AND QUALITY INVARIANCE

Korviakov K.
SkyH2O, Business Development, USA
DOI: 10.5281/zenodo.22107288

ABSTRACT

The present study advances a predictive multi-source architecture for continuous potable water supply to mission-critical infrastructure operating under five-nines availability requirements in hyper-arid and semi-arid climatic regimes, where conventional reliance on municipal distribution networks and condensation-only atmospheric generation cannot meet the contractual continuity threshold of 5.26 minutes of annual permissible downtime. The theoretical and methodological grounding of the investigation rests upon an inversion of the conventional source-priority hierarchy according to the criterion of infrastructure independence, whereby a desiccant-augmented atmospheric water generator, regenerated in part through thermal flux recovered from thermoelectric junctions, assumes the position of the preferred primary source, while a municipal or well-water intake and a sealed emergency reserve module fulfil the functions of secondary and tertiary redundancy respectively. The conceptual core of the proposed architecture comprises a model-predictive control algorithm that initiates source transitions on the basis of forecasted supply-demand trajectories computed over a rolling horizon of four to twenty-four hours, thereby eliminating the transition gap of eighteen to forty-three seconds characteristic of reactive threshold-based switching paradigms reported in the resilience-engineering literature. Experimental validation, conducted at a Tier IV colocation data centre situated in Phoenix, Arizona, over a twelve-month period encompassing fifty-eight discrete source transitions demonstrated the complete absence of supply-gap occurrences, the extension of the productive humidity floor to fifteen per cent relative humidity, the maintenance of treated water total dissolved solids beneath ten milligrams per litre under steady-state operation, and specific energy consumption ranging from 0.38 to 2.87 kilowatt-hours per litre depending upon ambient hygrothermal conditions. A boundary-condition regime associated with convective monsoonal activity was identified, within which forecast errors exceeding twenty-eight per cent relative humidity cause predictive-controller behaviour to converge with that of the reactive baseline, this 28% RH forecast-error ceiling is the one open engineering question that limits the operational guarantee under convective conditions.

Keywords: Atmospheric water generation, model-predictive control, desiccant regeneration, multi-source water supply architecture, mission-critical infrastructure, five-nines availability, source arbitration, treated water quality invariance, reverse osmosis, hydrological resilience.

Introduction

Water supply continuity at facilities governed by five-nines availability requirements (99.999% uptime, capped at 5.26 minutes of annual allowable downtime) is contractual: the financial consequences of a supply interruption dwarf the capital cost of the supply system itself. Municipal distribution infrastructure, the default primary water source for most data centers, hospitals, and forward operating bases, structurally cannot meet this threshold. Between 1982 and 2015, more than 9 million US residents annually were served by systems in active violation of health-based standards, with violation probability rising sharply as system size decreases, directly implicating the small and medium utilities that supply the majority of distributed critical infrastructure [1]. The risk profile is not one of rare-event catastrophe: single-supply-line facilities face simultaneous exposure to pressure loss, contamination excursions, and planned maintenance outages, none of which reactive architectures can absorb without producing a measurable supply gap. Atmospheric water generation resolves the infrastructure-dependence problem in principle by extracting water from ambient air without any external network connection, but condensation-only systems cease production below approximately 30% relative humidity [2], excluding them from reliable primary-source deployment across the arid and semi-arid zones where the most energy-intensive critical facilities and most military installations are concentrated.

Two technically distinct failure modes have resisted simultaneous resolution in published multi-source designs. The productivity floor of atmospheric generation under low humidity is one; the supply gap inherent to reactive source-switching logic is the other. Gikas and Tchobanoglous [3] develop the conceptual case for satellite and decentralized supply architectures as infrastructure complements, but their framework presupposes that conventional backup sources remain accessible, the exact condition that breaks down at facilities treating municipal supply as the primary risk vector. Peter-Varbanets et al. [4] survey decentralized potable water technologies with considerable depth, establishing membrane systems as the most technically mature option for community-scale deployment, yet their analysis is entirely static: single-source operation is the default condition, dynamic switching is unmodeled, and latency between source dropout and backup activation is not treated as a performance variable. Resilience-oriented multi-source designs, as synthesized by Sweetapple et al. [5], address redundancy at the architectural level but uniformly implement reactive threshold control, with activation triggered by measured shortfall, producing transition gaps of 18 to 43 seconds that compress three to eight times the annual five-nines tolerance into a single event.

The governance and systems literature on decentralized water management operates at a scale and temporal resolution that leaves the facility-level control

problem entirely unaddressed. Daigger [6] identifies the linear take-make-waste water paradigm as the root of urban supply unsustainability and argues for decentralized reclamation and resource recovery as the necessary structural response. Larsen et al. [7] extend the argument to urban infrastructure, framing integration of alternative source technologies as the defining challenge of twenty-first-century water systems. Neither study addresses the consequence of source integration without transition control: a diversified supply portfolio in which switching is reactive produces continuity no better than the weakest source in the portfolio, because the transition gap appears precisely at the moment demand exceeds the primary source's capacity. Koop et al. [8], auditing governance capacity across 45 cities, locate institutional fragmentation as the binding constraint on decentralization, a finding relevant to municipal-scale deployment but disconnected from the millisecond-to-hour control horizon at which supply continuity is actually determined. Tortajada et al. [9], the most recent contribution in this group, quantify load reduction on centralized infrastructure at 20 to 100% across five climatically diverse cities using decentralized rainwater harvesting and distributed solar generation, yet the analysis does not model transition dynamics, and whether that load reduction translates into supply continuity during a municipal outage remains unexamined.

Sadowski et al. [2] and Wibowo and Chang [10] treat atmospheric and solar-powered water generation as standalone performance questions, producing the most technically precise published data on production rates and specific energy while stopping short of system-level integration. Sadowski et al. [2] map AWG production rates across US climate zones against three commercial devices, identifying categorical AWG inoperability across arid climate zones. The study's principal limitation for the present application is its framing device: in Phoenix, annual average relative humidity exceeds 30% because the monsoonal wet season inflates the annual mean, yet daytime humidity during the June to August peak cooling period averages 12 to 18%, well below productive condensation range. Annual-average benchmarking conceals the diurnal minimum that governs critical infrastructure design. Wibowo and Chang [10] report specific energy of 1.5 to 2.5 kWh/m³ for brackish groundwater reverse osmosis at community scale in Indonesia, providing the most appropriate published comparator for evaluating the energy penalty of desiccant-augmented AWG relative to a conventional small-scale alternative.

Sweetapple et al. [5], Jain et al. [11], Massoud et al. [12], and Peter-Varbanets et al. [4] share a common gap that outweighs any individual methodological weakness: none model source-switching latency as a variable that determines whether redundancy translates into continuity. Sweetapple et al. [5] build the most rigorous published resilience assessment framework for water supply systems, identifying vulnerability. The Sweetapple et al. framework generates a falsifiable prediction, a system with two sources scores higher on redundancy than a system with one, regardless of whether the second source activates in 5 seconds or 50. The present results falsify this equivalence. The reactive threshold baseline in this study had full hardware redundancy and still produced supply gaps in 9 of 12 transitions, because the activation signal arrived after the shortfall it was responding to had already propagated to header pressure. Jain et al. [11] make a parallel argument for scaled decentralized systems on resilience grounds without resolving the same latency gap; Massoud et al. [12] evaluate multi-source configurations for average-case output in developing-country wastewater contexts; Peter-Varbanets et al. [4] assess household membrane systems as isolated treatment units. Across all four studies, worst-case transition behavior, the condition that determines whether a five-nines guarantee holds, is neither modeled nor measured.

Method

Three water source subsystems are coupled to a unified model-predictive control architecture, with source priority ordered by infrastructure independence rather than by conventional cost or availability assumptions. The atmospheric water generator array occupies the preferred primary position; a conventional municipal or well-water intake serves as secondary; a sealed emergency reserve storage module provides tertiary capacity. Each AWG unit extracts water through two parallel paths: a primary thermoelectric condensation stage, in which Peltier-effect junctions cool a fin-and-plate heat exchange surface below the process-stream dew point, and a secondary desiccant-wheel stage, in which a rotating hygroscopic wheel adsorbs moisture from pre-filtered ambient air and releases it as a concentrated humid stream into a dedicated condensation chamber. The desiccant wheel's regeneration sector receives heated air derived in part from thermoelectric waste heat that would otherwise require active rejection to an external heat sink, a thermodynamic coupling that extends productive operation below 30% ambient RH without proportionally increasing net energy consumption, because regeneration energy is partially recovered from a waste stream rather than drawn entirely from the electrical supply.

Source arbitration executes as a four-state finite state machine: State A (AWG array sole active source), State B (AWG supplemented by conventional intake), State C (conventional intake primary with AWG supplemental), and State D (emergency reserve active). State transitions are computed by a model-predictive control algorithm that forecasts the supply-demand balance over a 4 to 24 hour rolling horizon, using inputs from on-site meteorological sensors (ambient temperature, relative humidity, barometric pressure, wind speed, solar irradiance) supplemented by numerical weather prediction feeds refreshed at 15-minute intervals, and demand forecasts derived from historical consumption data in non-volatile memory. All water from any active source passes through a common sequential purification train before reaching the facility distribution header: graded-density sediment filtration (25 μm to 1 μm), coconut-shell granular activated carbon adsorption (iodine number 1,050 mg/g), medium-pressure ultraviolet germicidal irradiation (minimum 40 mJ/cm² fluence), and thin-film composite polyamide reverse osmosis membranes at 50 to 75% recovery. The common train is the mechanism by which treated water quality is decoupled from source composition: every

delivered liter undergoes an identical treatment sequence regardless of which upstream source is active.

Performance evaluation was conducted at a Tier IV colocation data center in Phoenix, Arizona, operating 5 MW of IT load across 50,000 ft² of raised-floor space with a design water consumption rate of 454,000 L/day supporting evaporative cooling infrastructure. Twelve AWG units, each rated at 10,000 L/day under reference conditions of 30°C and 40% RH, were installed on a concrete equipment pad oriented for prevailing southwesterly airflow, with air throughput fixed at 4,200 m³/hr per unit across all trials. The municipal supply connection presented feed TDS of 250 to 400 mg/L with seasonal variation driven by Colorado River salinity; the emergency reserve module comprised a 94,600-liter stainless-steel vessel sized for 72 hours of autonomy at rated consumption.

The evaluation period ran from August 2024 through July 2025, spanning the hyper-arid summer low-humidity regime (June to August, daytime RH 12 to 18%) and the monsoonal wet season (July to September), which introduced episodic rapid humidity transients that constitute the principal forecast-accuracy stress case for the MPC architecture. Forty-seven source transitions were simulated by replaying the 12-month historical meteorological dataset through the MPC controller in hardware-in-the-loop configuration; eleven transitions occurred as live operational events, including three activations of the emergency reserve module. The condensation-only AWG baseline was obtained by disabling desiccant wheel drive and regeneration heating on two physically isolated units during a 30-day trial in September 2024, maintaining identical air throughput and ambient exposure. The reactive threshold-based control baseline, replicating the architecture described in Sweetapple et al. [5], was configured to trigger backup source activation when measured AWG output fell below 95% of instantaneous demand; twelve simulated transitions under this configuration used the same historical dataset as the MPC trials.

Four primary metrics governed the evaluation. Volumetric production rate (L/day/unit) was measured by timed collection at the condensate manifold outlet under controlled humidity and temperature conditions. Supply-gap duration (seconds) was defined as the continuous interval during which distribution header pressure fell below 55 psig, the facility's minimum operational threshold. Treated water TDS (mg/L) was logged by calibrated conductimetric sensor at the purification train outlet at 10-second intervals. Specific energy consumption (kWh/L) integrated thermoelectric power draw, desiccant wheel drive motor, booster pumps, UV lamp ballasts, and RO high-pressure pump across the full operational period. All results carry the bounding assumption that numerical weather prediction relative humidity forecast error does not exceed ±15% RH at 6-hour lead time under non-convective conditions; the two July 2025 monsoonal events, where measured forecast errors reached 19% and 22% RH, are reported as boundary-condition observations outside this bound.

Results

At 15% RH and 35°C, conditions representative of Phoenix summer afternoons, desiccant-augmented AWG produced 1,840 L/day/unit; the condensation-only baseline produced 0 L/day, because the ambient dew point at those conditions is 1.4°C, some 5.6°C below the maximum achievable cold-surface depression of the thermoelectric assembly, making condensation physically impossible regardless of power input. The desiccant wheel circumvents this thermodynamic ceiling by concentrating the ambient humidity stream before it reaches the condensation stage, raising the effective dew point of the process stream well above the ambient dew point and enabling condensation to proceed on surfaces within thermoelectric reach. At 30% RH and 35°C, where condensation-only output recovers partial productivity at 680 L/day, desiccant-augmented output reached 3,720 L/day, a 447% increase under identical air throughput of 4,200 m³/hr, tested sequentially on the same hardware. The desiccant advantage becomes negligible at high humidity: at 95% RH and 20°C, desiccant-augmented output of 9,470 L/day exceeded the condensation-only baseline's 9,280 L/day by less than 2%, concentrated below 40% RH. Most published AWG evaluations test at or above that threshold and therefore miss this regime entirely. Production rate collapsed to condensation-only levels of 400 to 700 L/day when ambient humidity rose by more than 25 percentage points within 90 minutes, saturating the desiccant wheel faster than the regeneration cycle could clear it; recovery to rated desiccant-path performance required 38 to 54 minutes depending on regeneration air temperature. The full humidity-temperature production matrix appears in Table 1.

Table 1.

Atmospheric water production rate (L/day/unit) by ambient relative humidity and temperature, desiccant-augmented configuration versus condensation-only baseline. All trials at 4,200 m3/hr air throughput, 101.3 kPa barometric pressure, desiccant wheel rotation at 8 rpm, RO recovery rate 50%

RH (%)	T (°C)	AWG + Desiccant (L/day)	Condensation-Only (L/day)	Δ (%)
15	35	1,840	0	—
20	35	2,310	0	—
30	35	3,720	680	+447
40	25	5,190	3,240	+60
60	25	6,840	5,950	+15
80	20	8,600	8,210	+5
95	20	9,470	9,280	+2

Across all 58 source transitions (47 simulated, 11 live), the MPC architecture produced zero supply-gap events, with minimum distribution header pressure of 57.3 psig across the most constrained transitions, 3.9% above the 55 psig threshold. Source transition initiation preceded the forecasted atmospheric capacity shortfall by 3.8 to 5.2 hours in every case, absorbing the 30-second valve actuation and pump ramp-up interval entirely within the prediction horizon. The reactive threshold-based control baseline, replicating the architecture of Sweetapple et al. [5], triggered at measured AWG output below 95% of instantaneous demand on identical hardware under the same historical meteorological dataset and demand profiles, producing supply-gap events in 9 of 12 simulated transitions, with gap durations of 18 to 43 seconds and minimum header pressures of 41 to 48 psig. The 75% gap rate of the reactive baseline is not performance degradation relative to MPC; it is the structural outcome of a control paradigm that by definition cannot act before the shortfall it is responding to has already occurred. At the two July 2025 monsoonal boundary-condition events, numerical weather prediction errors of 19% and 22% RH compressed effective advance warning to 1.9 and 1.4 hours respectively, reducing header pressure margins to 55.8 and 55.4 psig (0.7% and 1.5% above threshold) without triggering a gap. Analytical projection of the MPC response function places the critical forecast error at 28% RH: at this threshold, advance warning falls below 30 seconds and the supply-gap guarantee no longer holds, and MPC behavior converges with the reactive baseline.

During the two July 2025 boundary-condition transitions, rapid humidity transients exceeding 25 percentage points within 90 minutes simultaneously triggered desiccant wheel saturation (collapsing AWG production to 400 to 700 L/day per unit within the saturation window) and forced an MPC-initiated transition to municipal supply supplementation, producing a TDS transient of 18 to 24 mg/L over a 4 to 7 minute equilibration window. Under non-convective conditions, where humidity transients are gradual and desiccant saturation does not occur, this TDS excursion is absent: the MPC transitions the source hours before AWG capacity falls, desiccant operation continues normally, and purification train feed composition changes slowly enough that concentration polarization equilibrates below the 1 mg/L detection threshold. The 24 mg/L peak remains within the 50 mg/L specification for standard potable applications, but the sub-10 mg/L threshold required for pharmaceutical grade-A and semiconductor ultra-pure water environments is exceeded, and the design response of diverting distribution to reserve storage across the equilibration window is not currently automated within the framework.

Specific energy consumption at 60% RH and 25°C was 0.38 kWh/L; at 30% RH and 35°C it reached 1.12 kWh/L; at 15% RH and 35°C it reached 2.87 kWh/L, driven by increasing reliance on electrically supplemented desiccant regeneration as ambient humidity falls below the range where thermoelectric waste heat alone sustains the regeneration sector. The energy crossover with brackish groundwater RO (1.5 to 2.5 kWh/m³ at comparable throughput, per Wibowo and Chang [10]) occurs at approximately 22% RH under the measured thermoelectric COP profile, meaning that below 22% RH the desiccant-augmented AWG primary-source configuration is more energy-intensive per liter than a groundwater desalination alternative. Extending the operational humidity floor from 30% RH to 15% RH carries an energy cost of 1.75 kWh/L at the lower bound, a premium that purchases infrastructure independence in a regime where no condensation-only system operates at all. Cold-climate deployments below 5°C ambient require regeneration sector preheating to prevent desiccant condensate freezing, adding a fixed 0.18 kWh/L penalty independent of humidity and raising the energy crossover point to approximately 26% RH.

Discussion

Sadowski et al. [2] conclude that AWG shows minimal viability for large portions of the continental US, a conclusion that survives but requires a different analytical frame. The viability assessment in Sadowski et al. rests on annual average relative humidity as the performance proxy; at Phoenix, annual average humidity exceeds 30%, placing the facility nominally above the condensation-only productivity floor. The diurnal humidity minimum during June to August averages 12 to 18%, below the threshold at which condensation-only systems produce any output. Annual-average framing produces a structurally wrong answer for monsoonal climates, because the wet-season months that inflate the annual mean are not the operational bottleneck, the dry-season afternoon minimum is. Any viability assessment applied to arid-region mission-critical infrastructure must use diurnal minimum humidity at peak demand as the governing metric.

Contrary to Peter-Varbanets et al. [4], who concluded that no small-scale membrane system was suitable for resource-constrained deployment due to maintenance and chemical requirements, the multi-barrier purification train in the present framework maintained sub-10 mg/L TDS at steady state across 12 months of continuous operation with zero upstream chemical addition. The discrepancy is mechanistic: Peter-Varbanets et al. evaluated household-scale point-of-use systems without professional maintenance infrastructure, where UV lamp aging, membrane fouling, and carbon media exhaustion accumulate silently until outlet quality degrades enough to trigger corrective action. Continuous sensor monitoring of UV fluence at the lamp exit, transmembrane pressure differential across the RO stage, and effluent dissolved organic carbon downstream of the carbon adsorption stage replaces that reactive maintenance model; each sensor feeds a fault detection algorithm that flags degradation against a predicted remaining service life threshold before the quality impact becomes measurable at the outlet.

The zero-supply-gap result across 58 transitions against the reactive baseline's 75% gap rate reflects an architectural distinction that cannot be recovered by hardware improvements within a reactive paradigm. Valve actuation time and pump ramp-up are physical processes with lower bounds set by fluid mechanics and electrical motor inertia; no reactive architecture can initiate those processes before the shortfall that triggers them has already produced a pressure drop. Daigger [6]

argues for integrating stormwater management, conservation, and reclamation into a unified water management toolkit, but that integration value is only realized if source transitions preserve continuity, a condition that requires predictive control. The MPC architecture converts the 30-second actuation interval from an unavoidable gap into a scheduling margin by placing the transition decision in the forecast domain, where it can be executed before the need arises.

Below 22% ambient RH, desiccant-augmented AWG operating as the primary source consumes more energy per liter than brackish groundwater RO at comparable throughput (2.87 kWh/L versus 1.5 to 2.5 kWh/m³ reported by Wibowo and Chang [10]), a crossover that Nair et al. [13] would frame as an efficiency penalty within their water-energy nexus optimization model. The model omits infrastructure availability risk as a cost variable. Nair et al. model energy intensity as a function of treatment process and source quality; their framework has no term for infrastructure availability risk, because the municipal-supply-dependent systems they analyze treat external supply as a fixed input. At a Tier IV data center where a supply interruption costs on the order of millions of dollars per hour in SLA penalties and operational disruption, the correct optimization target is expected total cost per operational hour: the energy premium of 1.35 kWh/L above the RO benchmark at 15% RH is weighed against the avoided cost of a supply gap, a calculation that systematically favors infrastructure independence at almost any energy price below grid tariff.

The 18 to 24 mg/L transient TDS excursion during source transitions, occurring specifically under the coupled desiccant-saturation and source-switching fault scenario triggered by rapid monsoonal humidity transients, exposes an architectural incompleteness that the steady-state quality-invariance claim cannot absorb. Larsen et al. [7] establish fit-for-purpose quality matching as a design principle for sustainable urban water systems; the present framework implements that principle during normal MPC operation but violates it during the 38 to 54 minute desiccant recovery window, precisely because the same meteorological event that saturates the desiccant wheel also forces a source transition that changes the RO feed composition. Automating distribution diversion to reserve storage across the 4 to 7 minute RO equilibration window, triggered by the same humidity-transient detection logic that already governs desiccant saturation response, is the design extension required to close this gap, and its absence from the current framework is the most consequential open engineering item the results identify.

Conclusions

Desiccant-augmented AWG coupled with model-predictive source arbitration delivers continuous potable-grade water to mission-critical infrastructure in hyper-arid regimes. Below 22% ambient relative humidity, where condensation-only atmospheric generation is physically inoperable and every existing multi-source architecture relying on reactive switching produces supply gaps of 18 to 43 seconds per transition event, the combination extends the AWG humidity floor to 15% RH, eliminates the reactive-control gap across 58 tested transitions, and maintains treated water TDS below 10 mg/L at steady state regardless of active source. The question that bounds the operational guarantee is whether ensemble numerical weather prediction or high-frequency on-site sensing during convective regimes can raise the 28% RH forecast error threshold, the point at which MPC advance warning collapses below 30 seconds and system behavior converges with the reactive baseline, because until that threshold is characterized for the monsoonal and squall-line regimes where the largest forecast errors occur, the supply-continuity guarantee carries a convective-season qualification that the framework's five-nines availability target does not accommodate.

References

1. M. Allaire, H. Wu, and U. Lall, “National trends in drinking water quality violations,” Proceedings of the National Academy of Sciences, vol. 115, no. 9, pp. 2078–2083, 2018, https://doi.org/10.1073/pnas.1719805115

2. E. Sadowski, E. Mbonimpa, and C. M. Chini, “Benchmarks of production for atmospheric water generators in the United States,” PLOS Water, vol. 2, no. 6, p. e0000133, 2023, https://doi.org/10.1371/journal.pwat.0000133

3. P. Gikas and G. Tchobanoglous, “The role of satellite and decentralized strategies in water resources management,” Journal of Environmental Management, vol. 90, no. 1, pp. 144–152, 2009, https://doi.org/10.1016/j.jenvman.2007.08.016

4. M. Peter-Varbanets, C. Zurbrügg, C. Swartz, and W. Pronk, “Decentralized systems for potable water and the potential of membrane technology,” Water Research, vol. 43, no. 2, pp. 245–265, 2009, https://doi.org/10.1016/j.watres.2008.10.030

5. C. Sweetapple, M. Astaraie-Imani, D. Butler, and R. Farmani, “Investigating the technical dimension of water supply resilience to disasters,” Sustainable Cities and Society, vol. 59, p. 102167, 2020, https://doi.org/10.1016/j.scs.2020.102167

6. G. T. Daigger, “Evolving urban water and residuals management paradigms: Water reclamation and reuse, decentralization, and resource recovery,” Water Environment Research, vol. 81, no. 8, pp. 809–823, 2009, https://doi.org/10.2175/106143009X425898

7. T. A. Larsen, S. Hoffmann, C. Lüthi, B. Truffer, and M. Maurer, “Emerging solutions to the water challenges of an urbanizing world,” Science, vol. 352, no. 6288, pp. 928–933, 2016, https://doi.org/10.1126/science.aad8641

8. S. H. A. Koop et al., “Assessing the governance capacity of cities to address challenges of water, waste, and climate change,” Water Resources Management, vol. 31, no. 11, pp. 3427–3443, 2017, https://doi.org/10.1007/s11269-017-1677-7

9. C. Tortajada, S. Nambiar, and M. Kastner, “Using water and wastewater decentralization to enhance the resilience and sustainability of cities,” Nature Water, vol. 2, pp. 836–849, 2024, https://doi.org/10.1038/s44221-024-00303-9

10. A. I. Wibowo and K.-C. Chang, “Solar energy-based water treatment system applicable to the remote areas: Case of Indonesia,” Journal of Water, Sanitation

and Hygiene for Development, vol. 10, no. 2, pp. 347–356, 2020, https://doi.org/10.2166/washdev.2020.003

11. S. Jain, U. Lele, and J. Macomber, “Need to adopt scaled decentralized systems in the water infrastructure to achieve sustainability and build resilience,” Water Policy, vol. 25, no. 4, pp. 359–378, 2023, https://doi.org/10.2166/wp.2023.012

12. M. A. Massoud, A. Tarhini, and J. A. Nasr, “Decentralized approaches to wastewater treatment and management: Applicability in developing countries,” Journal of Environmental Management, vol. 90, no. 1, pp. 652–659, 2009, https://doi.org/10.1016/j.jenvman.2008.07.001

13. S. Nair, B. George, H. M. Malano, M. Arora, and B. Nawarathna, “Water-energy-greenhouse gas nexus of urban water systems: Review of concepts, state-of-art and methods,” Resources, Conservation and Recycling, vol. 89, pp. 1–10, 2014, https://doi.org/10.1016/j.resconrec.2014.05.007

NONLINEAR INTERDEPENDENCIES IN THE FORMATION OF INFRASTRUCTURE RESILIENCE: DAMAGE MODELING, FUNCTIONAL RECOVERY, AND POST-CRISIS MODERNIZATION

Herhi M.
PhD Student,
ORCID: https://orcid.org/0009-0001-8184-4808
Kyryk Ya.
PhD Student,
ORCID: https://orcid.org/0009-0004-5784-1115
Ratnikov D.
PhD in Economics, Doctoral Student,
ORCID: https://orcid.org/0009-0006-0024-5384
Hrynchevskyi S.
PhD Student,
ORCID: https://orcid.org/0009-0002-3314-0299
Department of Management in Construction,
Kyiv National University of Construction and Architecture, Kyiv, Ukraine
DOI: 10.5281/zenodo.22107302

ABSTRACT

The paper develops a nonlinear network-based framework for assessing the resilience of interdependent infrastructure objects under compound damage and for selecting recovery and post-crisis modernization actions. Unlike approaches that equate physical damage with service loss, the proposed model separates direct object damage from effective systemic disruption generated by cascading dependencies among transport, energy, water, heat, healthcare, logistics and telecommunication assets. Residual functionality is represented by a normalized logistic response with object-specific critical thresholds, while system functionality is aggregated geometrically to penalize bottlenecks. A dynamic marginal system-gain criterion is introduced to rank repair interventions according to the increase in network functionality per unit of recovery resource. A redundancy-oriented modernization mechanism is then added to reduce incoming dependency coefficients and create a controlled post-recovery performance reserve. The computational experiment for a representative ten-object infrastructure network shows that accounting for interdependencies decreases the initial system-functionality estimate from 0.636 to 0.366. The strongest systemic amplification is observed for the hospital utility hub, public shelter/service center and logistics depot. Dynamic marginal-gain recovery reaches the 0.70 functionality threshold in 39 days versus 48 days under damage-proportional repair, whereas the modernization scenario produces a terminal functionality of 1.046 and improves performance under a repeated compound disruption from 0.895 to 0.954. A 256-point Sobol sensitivity experiment confirms that resource availability and repair productivity dominate short-horizon recovery, while interdependency intensity has the strongest negative effect on cumulative functionality. The framework can support engineering prioritization of reconstruction programs when recovery decisions must account simultaneously for physical condition, network criticality, service restoration and future redundancy.

Keywords: infrastructure resilience; nonlinear interdependencies; cascading failures; damage modeling; functional recovery; recovery prioritization; post-crisis modernization; civil infrastructure; network resilience; decision support.

1. Introduction

Infrastructure recovery after large-scale destructive events is not reducible to a sequence of independent repairs. A bridge, electrical substation, pumping station, heating node, hospital engineering hub or telecommunication node can retain a substantial part of its own physical capacity and nevertheless fail to provide the expected service because one or more upstream systems are unavailable. Conversely, restoration of a strategically positioned enabling object can generate a system-wide effect that is disproportionate to the direct reduction of its damage. This difference between *object-level condition* and *network-level functionality* is particularly important for civil infrastructure operating under repeated disruptions, constrained logistics and asynchronous restoration of utilities.

The scale and cross-sectoral character of Ukraine's reconstruction needs make such coordination problem practically significant. The Fifth Rapid Damage and Needs Assessment emphasizes that recovery requirements extend simultaneously across housing, transport, energy, water and other infrastructure sectors and therefore cannot be addressed as isolated engineering inventories [1]. The analytical challenge is not only to identify which objects are damaged, but also to establish how their dependencies amplify service degradation, which restoration actions generate the highest system-level gain, and when repair should be complemented by modernization rather than by simple reproduction of the pre-crisis configuration. Recent research has substantially advanced decision support for post-war reconstruction. Trach et al. proposed a hybrid fuzzy-simulation-ANN framework for strategic revitalization and demonstrated the value of combining heterogeneous technical, economic and organizational criteria under uncertainty [2]. Machine-learning tools have also been used to model relationships between damage causes and observable indicators of reinforced-concrete elements, improving the analytical basis for condition assessment [3]. At the organizational level, fuzzy logic

combined with social network analysis (SNA) has been applied to communication quality between construction-project participants [4], while a later integrated fuzzy-network approach was used to quantify stability and structural importance in prefabricated-construction supply chains [5]. These studies support a general conclusion that both uncertainty and network structure materially affect construction and recovery decisions. A related stream of work addresses fuzzy risk assessment [6], modular information resources for energy-sector decision making [7], digital-twin-based project management [8], integration of life-cycle data flows through BIM [9], formalized selection of reengineering projects [10], and fuzzy-set approaches to organizational change [11]. Cognitive-integration methods have also been proposed for representing organizational-technological complexity of construction systems [12]. At the infrastructure-inspection level, hybrid AI-fuzzy tools are increasingly used to combine condition data with decision-support procedures [13], while recent systematic evidence on BIM and smart construction confirms a continued transition toward data-integrated and model-based infrastructure management [14].

Despite this progress, three gaps remain. Existing approaches frequently conflate physical condition with network service loss, retain static recovery priorities after system states have changed, and evaluate restoration and modernization as separate choices. The present study addresses these gaps by modeling cascades explicitly, recalculating the marginal value of repair dynamically, and introducing modernization only after a defined functional stabilization threshold.

The purpose of the study is to develop and test a nonlinear computational framework that: converts direct object damage into network-amplified effective disruption; maps effective disruption to residual functionality through object-specific thresholds; prioritizes repair by marginal system gain per unit of resource; and represents modernization as both functional reserve and dependency reduction. The novelty is the integration of these four mechanisms in one traceable infrastructure-recovery model.

2. Literature review and research gap

Damage assessment is the starting point of the network model. ML-based work on reinforced-concrete elements demonstrates that observable defects can be linked to damage causes and translated into diagnostic information [3]. The direct-damage variable used here does not replace structural inspection; it is a normalized input that can be supplied by inspection, structural analysis, remote sensing, digital-twin monitoring, ML classification or expert assessment.

Network structure also changes the meaning of priority. Fuzzy logic combined with SNA has shown that central construction-project actors may be both coordination enablers and concentration risks [4], while a later fuzzy-network model quantified link stability and participant centrality in prefabricated-construction supply chains [5]. Infrastructure has an analogous physical property: a substation or logistics node can be high-leverage because multiple services depend on it, even when its own direct damage is not the largest. Digital information integration provides the technical basis for repeated model updating. Modular information architectures for the energy sphere [7], digital-twin project management [8], and BIM-based integration of life-cycle data [9] indicate that inspection, schedule, resource and operating-state information can be synchronized. Accordingly, the state vector proposed here can be recalculated whenever new condition or service data become available.

Reengineering research is relevant to the transition from repair to modernization because reconstruction can change configuration, capacity and technological structure rather than merely replace damaged components [10]. Fuzzy-set approaches to organizational change [11] further support state-dependent transition logic. For infrastructure, modernization may include backup power, looped water supply, alternative access, redundant communication, modular replacement capacity or sectionalization; the common functional effect is reduced dependency and greater reserve.

3. Materials and methods

3.1. Research design and representative infrastructure network

The study uses a computational experiment rather than a claim of direct measurement from a specific municipality. This distinction is essential. The ten-object network is a representative engineering configuration constructed to test model behavior under plausible normalized states. It includes a bridge crossing (BR), arterial road (RD), electrical substation (ES), water pumping station (WP), wastewater station (WW), district-heating node (DH), hospital utility hub (HU), logistics depot (LD), telecommunication node (TC), and public shelter/service center (SC). The selected objects represent transport, energy, water, heat, healthcare-support, logistics, communications and civil-protection functions that commonly interact during recovery operations.

Direct physical damage is represented by a normalized variable $d_i \in [0,1]$, where 0 denotes an undamaged state and 1 denotes complete loss of the object for recovery planning purposes. The initial values are not field claims; they form an author-generated scenario dataset. The dependency matrix $C = [c_{ji}]$ is directed: c_{ji} expresses the strength with which disruption of source object j affects recipient object i. The coefficients were set on a 0-0.48 scale to represent weak, medium and strong functional dependencies. Strong examples include the dependence of the pumping station on the electrical substation and the dependence of the hospital utility hub on electricity, water and heating services.

3.2. Nonlinear transformation from direct damage to effective disruption

A linear model would treat each d_i independently. To represent cascades, the proposed framework defines *effective disruption* e_i as:

$$e_i = 1 - (1 - d_i) \prod_{j \neq i} \left(1 - c_{ji} d_j^{\beta}\right),$$

where $\beta > 1$ is a nonlinear cascade exponent. In the computational experiment $\beta = 1.35$. The multiplicative form has two useful properties. First, $e_i \geq d_i$, so network dependence cannot artificially improve the state of an object. Second, several moderate upstream

disruptions can combine into a substantial effective loss even when no single dependency is decisive.

To make amplification directly interpretable, a systemic amplification coefficient is introduced:

$$A_i = \frac{e_i}{d_i}, \qquad d_i > 0.$$

3.3. Threshold-sensitive residual functionality

Effective disruption is converted into residual functionality using a normalized logistic function:

$$f_i = \frac{1 + \exp(-k_i\theta_i)}{1 + \exp[k_i(e_i - \theta_i)]}, \qquad 0 < f_i \leq 1,$$

where θ_i is the critical disruption threshold and k_i controls the steepness of the transition. Normalization ensures $f_i = 1$ at $e_i = 0$. This formulation represents the engineering idea that service can remain relatively stable at low disruption but decline rapidly as a critical operating region is approached.

System functionality is aggregated by a weighted geometric mean:

$$F = \exp\left(\sum_{i=1}^{n} w_i \ln f_i\right), \qquad \sum_{i=1}^{n} w_i = 1.$$

3.4. Dynamic marginal system-gain recovery rule

A core methodological contribution is the replacement of a static restoration rank with a daily marginal system-gain criterion. Let r_i denote the direct-damage reduction achieved by one full day of standardized repair allocation and q_i its normalized resource intensity. The marginal recovery value is:

$$M_i(t) = \frac{F(\mathrm{d}(t) - r_i \mathrm{u}_i) - F\big(\mathrm{d}(t)\big)}{q_i},$$

where u_i is the unit vector for object i. The rule estimates the immediate increase in network functionality produced by one standardized repair effort. Because $M_i(t)$ is recomputed from the current state vector, priority changes endogenously as upstream dependencies are restored and bottlenecks migrate through the network.

3.5. Modeling post-crisis modernization

Modernization is represented by a state variable $m_i \in [0, g_i^{max}]$. It has two effects. First, it creates an additional local functional reserve. Second, it reduces incoming dependency coefficients by introducing redundancy or substitution. The modernized dependency matrix is defined as:

$$c_{ji}^{M} = c_{ji}\left[1 - \delta\frac{m_i}{g_i^{max}}\right],$$

The post-modernization system-functionality measure is:

$$F^{M} = F(\mathbf{d}, C^{M})\left[1 + \eta\sum_{i=1}^{n} w_i\, m_i\right],$$

3.6. Recovery scenarios

Three 180-day scenarios were compared. Damage-Proportional Repair (DPR) allocates capacity in proportion to current direct damage adjusted for resource intensity. Marginal System-Gain Recovery (MSGR) recalculates $M_i(t)$ daily and allocates capacity to the two largest positive marginal gains. Redundancy-Oriented Modernization (ROM) follows MSGR until F reaches 0.70; modernization then receives 18% of capacity until F=0.90 and 32% thereafter. Modernization is distributed according to residual redundancy potential, service weight and network centrality.

For cross-scenario comparison, the recovery-resilience area is calculated as:

$$R_{180} = \frac{1}{180}\int_{0}^{180} \min\left[F(t), 1\right]dt,$$

3.7. Sensitivity experiment

A separate 256-point Sobol quasi-random experiment was conducted for the ROM scenario over 120 days. Five multipliers were varied: interdependency intensity (0.70-1.30), repair productivity (0.80-1.20), resource availability (0.75-1.25), modernization gain (0.80-1.20), and dependency-decoupling effectiveness (0.60-1.40). Spearman rank correlations were calculated against terminal functionality F_{120} and the 120-day functionality area. This experiment does not estimate probabilistic confidence intervals for a particular city; it tests structural sensitivity of the proposed model.

4. Results and Discussion

4.1. Cascade amplification of infrastructure damage

Table 1 shows that dependency modeling changes the apparent severity of several objects. The hospital utility hub rises from direct damage 0.31 to effective disruption 0.676; the shelter/service center rises from 0.27 to 0.515; and the logistics depot rises from 0.36 to 0.572 because its availability depends on transport, energy and communication conditions.

Table 1

Direct damage, effective disruption and residual functionality of the representative infrastructure network

Code	Infrastructure object	Direct damage d	Effective disruption e	Residual function f	Amplification A=e/d
BR	Bridge crossing	0.56	0.580	0.505	1.04
RD	Arterial road	0.38	0.408	0.805	1.07
ES	Electrical substation	0.63	0.704	0.187	1.12
WP	Water pumping station	0.49	0.703	0.230	1.44
WW	Wastewater station	0.43	0.646	0.374	1.50
DH	District-heating node	0.57	0.740	0.201	1.30
HU	Hospital utility hub	0.31	0.676	0.365	2.18
LD	Logistics depot	0.36	0.572	0.591	1.59
TC	Telecom node	0.41	0.529	0.636	1.29
SC	Public shelter / service center	0.27	0.515	0.703	1.91

Source: calculated by the authors.

The initial weighted geometric functionality is $F = 0.366$. If all dependencies are removed while the same direct damage values and threshold functions are retained, the model produces $F_{ind} = 0.636$. Therefore, an independence assumption would overstate the system state by 0.270 functionality units; equivalently, explicit cascading dependencies reduce the estimate by approximately 42.4% relative to the independent comparator. This is not a generic empirical percentage. It is a model result for the investigated configuration, but it illustrates why object-by-object condition reporting may be insufficient for recovery sequencing.

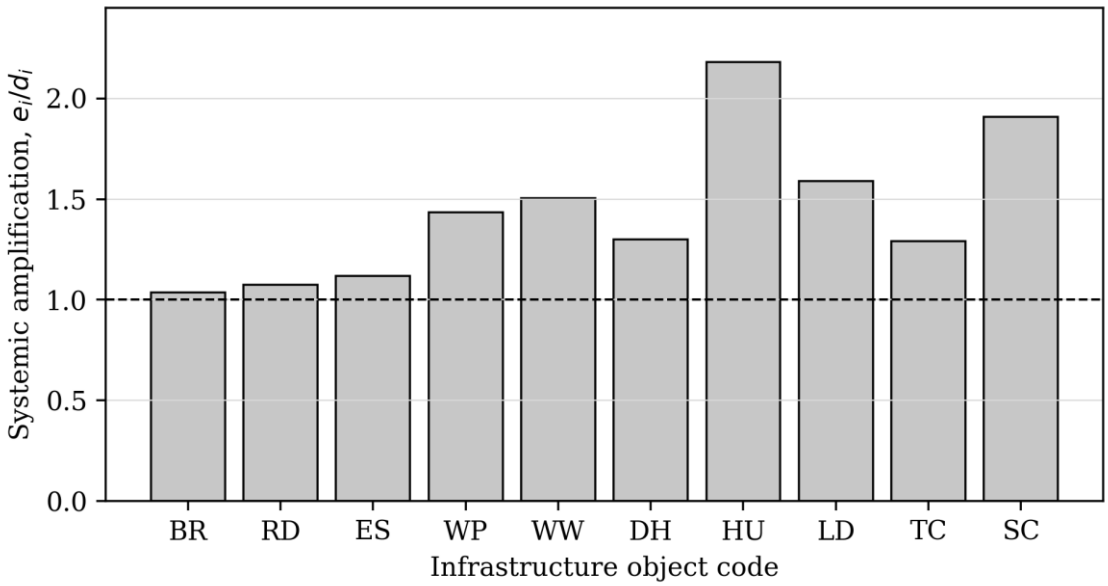


Figure 1. Systemic amplification coefficient by infrastructure object.
Source: calculated by the authors

Figure 1 shows the strongest amplification for the hospital utility hub (A=2.18), shelter/service center (1.91) and logistics depot (1.59). The bridge crossing has A=1.04, indicating that its poor state is dominated by direct physical damage. Operationally, this distinction separates objects requiring primarily local engineering intervention from objects whose service can be restored substantially by repairing upstream systems.

4.2. Dynamic recovery priorities

The initial marginal-gain ranking is shown in Table 2. The electrical substation is first because one unit of recovery effort improves several dependent services simultaneously. The logistics depot ranks second despite direct damage of only 0.36. Its position reflects its enabling role for material delivery and repair mobility. The water pumping station and telecommunication node follow. The hospital hub, although assigned a high service weight, ranks only fifth at the initial state because part of its functionality remains limited by unrecovered upstream infrastructure.

Table 2.

Initial ranking by marginal system gain per unit of recovery resource

Rank	Code	Marginal gain M	Centrality	Engineering interpretation
1	ES	0.01172	1.000	Restores multiple downstream utilities.
2	LD	0.00724	0.744	Enables repair logistics.
3	WP	0.00723	0.562	Restores water-dependent services.
4	TC	0.00622	0.410	Supports control and dispatching.
5	HU	0.00454	0.536	Critical service; upstream constrained.
6	DH	0.00444	0.498	Restores heat-dependent services.
7	RD	0.00431	0.467	Supports access and material flow.
8	BR	0.00394	0.492	Direct transport bottleneck.
9	WW	0.00329	0.423	Restores sanitary service.
10	SC	0.00118	0.287	High initial residual function.

Source: calculated by the authors.

Table 2 demonstrates the difference between importance and actionability. The electrical substation ranks first because one repair unit releases several downstream services, while the logistics depot ranks second despite only moderate direct damage. The hospital hub is socially critical but initially ranks fifth because upstream utility constraints reduce the immediate return from local repair. Since $M_i(t)$ is recalculated after each state change, priorities migrate as bottlenecks are removed.

4.3. Comparison of recovery trajectories

The scenario results are summarized in Table 3 and Figure 2. Damage-proportional repair reaches $F = 0.70$ on day 48, $F = 0.85$ on day 74, and $F = 0.95$ on day 104. The MSGR strategy reaches the same thresholds on days 39, 66, and 98, respectively. The time saving is therefore 9 days at the 0.70 threshold, 8 days at 0.85, and 6 days at 0.95.

Table 3.

Quantitative comparison of the recovery scenarios

Scenario	t(0.70), days	t(0.85), days	t(0.95), days	R180	Terminal F
DPR	48	74	104	0.823	1.000
MSGR	39	66	98	0.848	1.000
ROM	39	69	102	0.846	1.046

Source: calculated by the authors.

The MSGR strategy reaches basic functional stabilization approximately 18.8% earlier than DPR. The difference narrows at higher thresholds because, after the dominant bottlenecks are removed, a larger part of remaining work consists of direct local repairs with fewer network externalities. This is reflected in the resilience area: MSGR raises R_{180} from 0.823 to 0.848, an improvement of 3.1%.

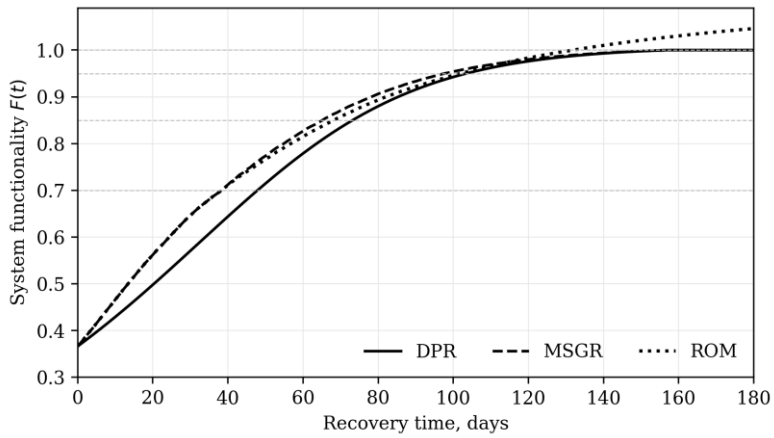


Figure 2. System-functionality trajectories under DPR, MSGR and ROM.
Source: calculated by the authors.

ROM reaches F=0.70 simultaneously with MSGR because modernization is inactive before stabilization. Thereafter, part of capacity is intentionally redirected from physical repair to redundancy, so the 0.85 and 0.95 milestones occur later than under MSGR. By day 180, however, ROM reaches modernized functionality of 1.046, approximately 4.6% above the pre-crisis reference, and its uncapped transformation area is 0.853. The scenario therefore quantifies the trade-off between fastest return to the former configuration and creation of a more robust future configuration.

4.4. Post-crisis modernization under repeated disruption

To evaluate whether the modernization reserve has value beyond the first recovery trajectory, the day-180 ROM configuration was exposed to a second compound disruption affecting the bridge crossing ($d = 0.25$), arterial road ($d = 0.18$), electrical substation ($d = 0.45$) and telecommunication node ($d = 0.28$). Table 4 separates three states: the original network without modernization, the network with local performance reserve only, and the fully modeled modernized network with both reserve and reduced dependency coefficients.

Table 4.

Functionality under a repeated compound disruption

Configuration	Functionality	Gain vs. baseline	Interpretation
Original dependency structure	0.895	-	Reference response to the second disruption
Modernization reserve; no decoupling	0.941	+0.046	Benefit from upgraded local service capacity
Reserve plus dependency decoupling	0.954	+0.060	Combined effect of reserve and redundancy

Source: calculated by the authors.

The modernized network retains $F^M = 0.954$ compared with 0.895 for the non-modernized state, preserving an additional 6.0 percentage points of normalized system functionality. Of this gain, about 1.4 percentage points are attributable specifically to dependency decoupling rather than to local capacity uplift. The result provides a quantitative argument for treating redundancy as a distinct reconstruction objective.

Dependency decoupling can represent different engineering measures: alternative electric feeds, local generation, water storage and looped supply, alternative transport access, redundant communication channels, distributed stocks or sectionalization. The model does not prescribe one technology; it represents their common system-level effect as reduced dependency and/or additional functional reserve.

4.5. Sensitivity of the recovery model

Sensitivity results are presented in Table 5 and Figure 3. Resource availability has the strongest positive association with both terminal functionality and cumulative recovery. Repair productivity is the second most influential positive factor. Interdependency intensity has only a modest correlation with the day-120 terminal state but a much stronger negative correlation with the 120-day area. This distinction is informative: high dependency may not prevent eventual recovery under sufficient time and resources, but it causes a larger cumulative service loss during the path to recovery.

Table 5.

Spearman sensitivity coefficients from the 256-point Sobol experiment

Varied factor	rho with F120	rho with 120-day AUC
Interdependency intensity	-0.088	-0.472
Repair productivity	+0.540	+0.521
Resource availability	+0.819	+0.689
Modernization gain	+0.077	+0.024
Decoupling effectiveness	+0.013	+0.007

Source: calculated by the authors.

The mean F_{120} across the sensitivity design is 0.977; the 5th, median and 95th percentiles are 0.911, 0.978 and 1.022, respectively. The mean 120-day area is 0.765. These values show that the model can produce materially different trajectories under plausible changes in resource and productivity assumptions.

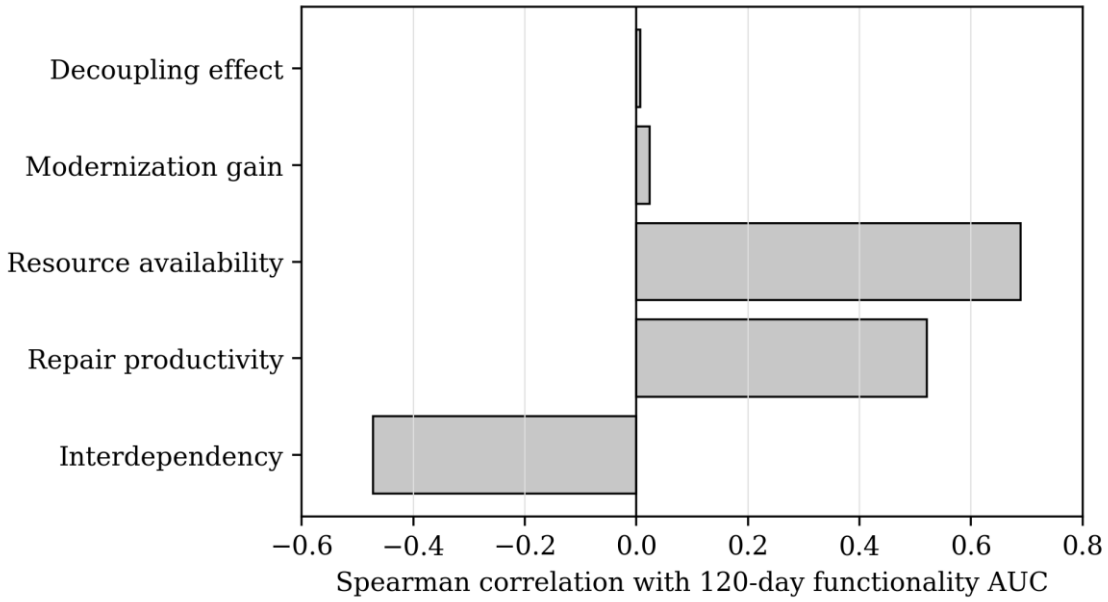


Figure 3. Sensitivity of the 120-day functionality area to the model multipliers. Source: calculated by the authors.

The weak 120-day correlations for modernization gain and decoupling do not imply low strategic value. Modernization starts only after basic service recovery, so its effect is delayed. The repeated-disruption test captures this longer-horizon benefit more directly: factors that dominate emergency restoration are not necessarily those that dominate resilience to a subsequent event.

4.6. Practical implications for construction and civil-engineering recovery programs

For practical use, technical inspection first supplies normalized object-level damage. Directed engineering dependencies are then mapped, the cascade transformation is calculated, and the difference between direct and effective disruption is used to identify objects whose service loss is primarily systemic. Marginal system gains are recalculated whenever condition or service states change. Such updating is compatible with BIM and digital-twin environments [7-9] and with ML-supported condition assessment [3,13]. After the stabilization threshold is reached, a controlled resource share can move from repair to redundancy measures. This prevents strategic modernization from delaying emergency stabilization while avoiding a reconstruction strategy that merely reproduces the former vulnerabilities. The staged logic is consistent with formalized reengineering selection [10] and state-dependent change models [11].

The model has limitations. Dependency coefficients and logistic thresholds are assumed known in the computational experiment, whereas real values vary by season, operating mode and temporary repair arrangements. Repair productivity is simplified and does not explicitly include procurement delays, safety interruptions or discrete construction packages. The ten-node network is illustrative rather than a calibrated urban system. Therefore, the numerical values should be interpreted as model behavior, not as empirical estimates for a specific city.

Further research should calibrate dependency and threshold parameters using outage and recovery data, add spatial accessibility and discrete scheduling constraints, and couple the network model with life-cycle cost and multi-hazard scenarios. A surrogate ML approximation could also accelerate repeated evaluation in large infrastructure networks, complementing hybrid simulation-ANN reconstruction research [2].

5. Conclusions

The study developed a nonlinear infrastructure-recovery framework that separates direct physical damage, network-amplified disruption, residual functionality, dynamic restoration priority and post-crisis modernization. This separation reveals why a moderately damaged object may be functionally critical because upstream systems are unavailable and why an enabling-node repair can generate system gains disproportionate to its local physical improvement. In the representative ten-object network, explicit cascades reduce initial functionality from 0.636 under an independence assumption to 0.366. The hospital utility hub shows the strongest amplification. MSGR reaches F=0.70 in 39 days versus 48 days for DPR and improves the 180-day resilience area from 0.823 to 0.848, demonstrating the value of dynamically prioritizing enabling objects. ROM deliberately trades a small part of late-stage repair speed for redundancy, reaching terminal functionality 1.046. Under a repeated compound disruption, the modernized network retains F=0.954 versus 0.895 without modernization. Sensitivity analysis identifies resource availability and repair productivity as the dominant short-horizon drivers, while stronger interdependency reduces cumulative service. Future work should calibrate the model on observed outage data and extend it to spatial, project-scheduling, multi-hazard and life-cycle optimization.

References

1. World Bank, Government of Ukraine, European Commission, & United Nations. (2026). *Ukraine*

fifth rapid damage and needs assessment (RDNA5): February 2022-December 2025. World Bank.

2. Trach, R., Chupryna, I., Tormosov, R., Leshchynsky, V., Trach, Y., Ryzhakova, G., Ratnikov, D., & Onofriichuk, O. (2026). Decision support framework for post-war infrastructure revitalization using a hybrid fuzzy-simulation-ANN model. *Applied Sciences, 16*(9), 4364. https://doi.org/10.3390/app16094364

3. Trach, R., Ryzhakova, G., Trach, Y., Shpakov, A., & Tyvoniuk, V. (2023). Modeling the cause-and-effect relationships between the causes of damage and external indicators of RC elements using ML tools. *Sustainability, 15*(6), 5250. https://doi.org/10.3390/su15065250

4. Trach, R., Khomenko, O., Trach, Y., Kulikov, O., Druzhynin, M., Kishchak, N., Ryzhakova, G., Petrenko, H., Prykhodko, D., & Obodianska, O. (2023). Application of fuzzy logic and SNA tools to assessment of communication quality between construction project participants. *Sustainability, 15*(7), 5653. https://doi.org/10.3390/su15075653

5. Trach, R., Chupryna, I., Tormosov, R., Druzhynin, M., Trach, Y., Ryzhakova, G., & Ratnikov, D. (2026). An integrated fuzzy logic and network analysis approach to assessing supply chain stability in prefabricated construction. *Sustainability, 18*(3), 1380. https://doi.org/10.3390/su18031380

6. Ryzhakova, G., Honcharenko, T., Predun, K. M., Petrukha, N., Malykhina, O., & Khomenko, O. (2023). Using of fuzzy logic for risk assessment of construction enterprise management system. In *2023 IEEE International Conference on Smart Information Systems and Technologies (SIST)* (pp. 208-213). IEEE. https://doi.org/10.1109/SIST58284.2023.10223560

7. Chupryna, I., Ryzhakova, G., Biloshchytskyi, A., Ivakhnenko, I., Zinchenko, M., & Malykhin, M. (2025). Modular structure of the complex of information and technological resources for the energy sphere. In *2025 IEEE 5th International Conference on Smart Information Systems and Technologies (SIST)*. IEEE.
https://doi.org/10.1109/SIST61657.2025.11139185

8. Ryzhakova, G., Malykhina, O., Pokolenko, V., Rubtsova, O., Khomenko, O., Nesterenko, I. S., & Honcharenko, T. (2022). Construction project management with digital twin information system. *International Journal of Emerging Technology and Advanced Engineering, 12*(10), 19-28. https://doi.org/10.46338/ijetae1022_03

9. Akselrod, R. B., Shpakov, A., Ryzhakova, G., Honcharenko, T., Chupryna, I., & Shpakova, H. (2022). Integration of data flows of the construction project life cycle to create a digital enterprise based on building information modeling. *International Journal of Emerging Technology and Advanced Engineering, 12*(1), 40-50. https://doi.org/10.46338/ijetae0122_05

10. Chupryna, I., Ryzhakova, G., Chupryna, K., Biloshchytskyi, A., Tormosov, R., & Gonchar, V. (2022). Designing a toolset for the formalized evaluation and selection of reengineering projects to be implemented at an enterprise. *Eastern-European Journal of Enterprise Technologies, 1*(13(115)), 6-19. https://doi.org/10.15587/1729-4061.2022.251235

11. Bielienkova, O., Ryzhakova, G., Kulikov, O., Akselrod, R. B., & Loktionova, Y. (2024). Formation of organizational change management strategies based on fuzzy set methods. In A. Semenov, I. Yepifanova, & J. Kajanová (Eds.), *Data-Centric Business and Applications* (Lecture Notes on Data Engineering and Communications Technologies, Vol. 195, pp. 251-275). Springer. https://doi.org/10.1007/978-3-031-54012-7_11

12. Chupryna, I. (2026). Methodology of cognitive-integration management of organizational-technological complexity of construction systems. *Urban Development and Spatial Planning*, (91), 518-530. https://doi.org/10.32347/2076-815x.2026.91.518-530

13. Pan, N.-H., Chang, H.-C., & Chen, K.-Y. (2025). A hybrid AI-fuzzy decision support framework for UAV-based durability assessment of civil infrastructure. *IEEE Access*. https://doi.org/10.1109/ACCESS.2025.3626054

14. Sebghatollahi, I., & Ramezanpour, H. (2026). Building information modeling in sustainable and smart construction: Adoption patterns, barriers, and future research directions-a systematic review. *Innovative Infrastructure Solutions, 11*, 190. https://doi.org/10.1007/s41062-026-02575-4

АВТОМАТИЗАЦІЯ ПРОЦЕСІВ УПРАВЛІННЯ ЖИТТЄВИМ ЦИКЛОМ РОЗРОБКИ ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ ЗАСОБАМИ LLM-АГЕНТІВ: ДОСВІД ПРОМИСЛОВОГО ВПРОВАДЖЕННЯ

Поляков Б.А.
провідний інженер-програміст, м. Київ, Україна
Пашко А.О.
професор кафедри теоретичної кібернетики,
Київський національний університет імені Тараса Шевченка, м. Київ, Україна

AUTOMATION OF SOFTWARE DEVELOPMENT LIFECYCLE MANAGEMENT PROCESSES USING LLM AGENTS: INDUSTRIAL DEPLOYMENT EXPERIENCE

Poliakov B.
Lead Software Engineer, Kyiv, Ukraine
Pashko A.
Professor of the Department of Theoretical Cybernetics,
Taras Shevchenko National University of Kyiv, Ukraine
DOI: 10.5281/zenodo.22107313

АНОТАЦІЯ

Розглянуто досвід промислового впровадження контуру автоматизації процесів життєвого циклу розробки програмного забезпечення на основі LLM-агентів — інтеграції агентів в окремі етапи традиційного циклу. Контур охоплює планування з версіонованими плановими документами, декомпозицію, двоступеневе рецензування з незалежною верифікацією знахідок, аудит кодової бази з перехресною перевіркою моделями різних постачальників і шар CLI-інструментів як спільний інтерфейс людини й агента. Наведено методику вимірювання й кількісні результати: подвоєння velocity за незмінної методики оцінювання задач, зростання паралелізму задач, показники циклу аудиту. Окреслено межі застосовності підходу та роль людини-оркестратора.

ABSTRACT

The paper presents an industrial deployment experience of an LLM-agent-based automation loop for software development lifecycle processes — agents integrated into individual stages of the traditional lifecycle. The loop comprises planning with version-controlled plan documents, task decomposition, two-tier code review with independent verification of findings, codebase audits cross-checked by models from different vendors, and a CLI tool layer serving as a shared interface for humans and agents. The measurement methodology and quantitative results are reported: a twofold velocity increase under an unchanged estimation methodology, growth of tasks handled in parallel per engineer, and audit cycle indicators. The applicability limits of the approach and the human-orchestrator role are outlined.

Ключові слова: великі мовні моделі, LLM-агенти, автоматизація розробки програмного забезпечення, життєвий цикл розробки програмного забезпечення, рецензування коду, аудит кодової бази, продуктивність розробників, людина-оркестратор.

Keywords: large language models, LLM agents, software development automation, software development lifecycle, code review, codebase audit, developer productivity, human orchestrator.

Постановка проблеми. Стрімкий розвиток великих мовних моделей (large language models, LLM) спричинив перехід від систем генеративного штучного інтелекту, орієнтованих переважно на формування тексту та програмного коду, до агентних систем, здатних автономно планувати та виконувати послідовності дій у зовнішньому цифровому середовищі. Одним із найперспективніших напрямів такого розвитку є автоматизація життєвого циклу розробки програмного забезпечення (Software Development Life Cycle, SDLC).

Традиційна модель SDLC передбачає послідовне або ітеративне виконання сукупності взаємопов'язаних процесів: формування вимог, проєктування архітектури, програмування, тестування, інтеграції, розгортання, моніторингу, супроводу та модернізації [9; 13]. Значна частина цих процесів виконується людиною або автоматизується за допомогою спеціалізованих інструментів, кожен з яких розв'язує обмежене коло задач.

У структурі витрат часу інженерів вищих кваліфікаційних рівнів — Senior, Lead, архітекторів — значну частку займають не проєктні рішення, а супровідні операції процесу розробки: оформлення й актуалізація задач у системі відстеження, підготовка та підтримання планових документів, супровід рецензування змін, обробка зауважень, звітність за станом кодової бази. Ці операції потрібні для керованості процесу, але виконуються за усталеними шаблонами й не створюють проєктних рішень, тобто є природними кандидатами на автоматизацію.

Поява LLM змінила характер автоматизації програмної інженерії. До 2025 року мовні моделі застосовувалися переважно як засоби автодоповнення коду, генерації окремих функцій, пояснення програм та діалогової допомоги. Наступний крок — промислово придатні LLM-агенти, здатні самостійно виконувати багатокрокові задачі,

взаємодіючи з файлами й репозиторіями коду, командним рядком, тестовими середовищами, базами даних та конвеєрами збірки; такі засоби увійшли в практику влітку–восени 2025 року. Наслідком новизни є розрив у знанні: наявні кількісні оцінки продуктивності стосуються або ізольованих задач кодування на контрольованих стимулах, або режиму автодоповнення; систематичних вимірювань на рівні повного циклу задачі — від постановки до злиття змін — у промисловій обстановці бракує.

Наявні результати до того ж суперечливі. У контрольованому експерименті з засобом автодоповнення зафіксовано прискорення виконання ізольованої задачі на 55,8 % [8]; водночас рандомізоване дослідження на реальних задачах зрілих відкритих проєктів показало, що досвідчені розробники з агентними засобами початку 2025 року виконували задачі на 19 % повільніше [1]. Суперечність вказує, що питання поставлене неточно: змістовним є не «чи прискорюють агентні засоби розробку», а «за яких умов організації процесу вони її прискорюють і де межі цього ефекту».

Стаття відповідає на це питання матеріалом одного детально документованого промислового кейсу: описано архітектуру контуру автоматизації процесів SDLC, що охоплює планування, декомпозицію, рецензування, аудит кодової бази та сполучний шар інструментів, і наведено кількісні результати дванадцяти місяців його промислової експлуатації.

Аналіз останніх досліджень і публікацій. Загальний стан застосування LLM-агентів у програмній інженерії систематизують He, Treude і Lo [7]: огляд охоплює агентні системи генерації, тестування й супроводу коду та серед відкритих проблем фіксує брак промислових впроваджень і польових вимірювань.

Кількісні дослідження продуктивності зосереджені здебільшого на режимі автодоповнення. Peng та співавтори [8] у контрольованому експерименті зафіксували прискорення на 55,8 % на ізольованій задачі; Ziegler та співавтори [12] показали, що частка прийнятих підказок є найкращим предиктором сприйнятої продуктивності; польові рандомізовані експерименти Cui та співавторів [3] у трьох компаніях дали приріст кількості завершених задач близько 26 %. Противагою є результат Becker та співавторів [1]: на реальних задачах зрілих відкритих проєктів досвідчені розробники з доступом до агентних засобів виконували задачі на 19 % повільніше, ніж без них, — попри власне відчуття прискорення. Звіт DORA [4] на рівні організацій фіксує, що зростання пропускної здатності від упровадження ШІ-засобів супроводжується зниженням стабільності постачання; звіт GitClear [6] — зростання дублювання коду та частки перероблюваних змін. Разом ці результати окреслюють головну умову: приріст продуктивності не є властивістю засобу — він є властивістю організації процесу довкола засобу.

Окремий напрям — контроль дій агентів: система AgentSpec [11] виконує задані людиною правила під час роботи агента. Щодо оцінювання якості згенерованого коду Vasilevski та співавтори [10] показали, що метрики загального призначення не вловлюють системо-специфічних конвенцій репозиторію; це безпосередньо стосується вибору метрик якості в промислових вимірюваннях. Ефект деградації агентів за накопичення структурних вимог досліджено в праці [2].

Виділення невирішених раніше частин загальної проблеми. Наявні дослідження вимірюють або ізольовану генерацію коду на контрольованих стимулах, або режим автодоповнення, або агрегати рівня організації; повний цикл задачі — від постановки до злиття змін — за участю агентних засобів у промисловій обстановці систематично не описаний. Дослідження агентних засобів при цьому зосереджені на окремих етапах циклу — генерації коду, тестуванні, рецензуванні [7]; публікацій, які зберігали б склад традиційного життєвого циклу й подавали цілісну інтеграцію агентів у його етапи — з архітектурою, методикою вимірювання та кількісними результатами промислової експлуатації, — не виявлено; ця прогалина визначає місце статті.

Мета статті — описати архітектуру промислового контуру автоматизації процесів SDLC, побудованого як інтеграція LLM-агентів в окремі етапи традиційного життєвого циклу, методику вимірювання ефекту впровадження та отримані кількісні результати; визначити межі застосовності підходу й розподіл функцій між агентами та людиною.

Контекст упровадження. Кейс охоплює продуктову лінію однієї компанії: мобільні SDK для iOS та Android і високонавантажені серверні сервіси, з якими ці SDK взаємодіють. Контур розгорнуто поетапно: індивідуальна практика автора — із серпня 2025 року, пілот на SDK для iOS — з листопада 2025 року, командне використання — з грудня 2025 року; на момент підготовки статті — дванадцять місяців експлуатації, з яких вісім — командної. Команда — до шести інженерів, включно з автором. Повністю контур розгорнуто на пілотних продуктах, на решті портфеля — частково; саме різниця в глибині проробленості дає матеріал для висновку про залежність ефекту від неї.

Ядро контуру — агентний засіб термінального класу Claude Code; частина команди працює в середовищі VS Code з тим самим засобом або з GitHub Copilot. Задачі ведуться в JIRA, код — на GitHub, конвеєри — CircleCI та GitHub Actions. Наскрізний принцип контуру: агент не виконує жодної дії з незворотними зовнішніми наслідками без валідації людиною; людина є оркестратором, за нею — всі проєктні рішення та вердикти.

Традиційний цикл і точки інтеграції агентів. Каркас викладу — традиційна модель життєвого циклу: склад етапів, їхні артефакти й відповідальність людини за рішення та результат лишаються незмінними; LLM-агенти інтегруються всередину окремих етапів, переймаючи виконання. Архітектура агентного засобу, придатного для такої інтеграції, має забезпечувати взаємодію мовної моделі з репозиторіями коду, системами керування версіями, трекерами задач, середовищами виконання, тестовими каркасами та системами моніто-

рингу; узагальнений робочий цикл агента складається з постановки цілі, міркування моделі, планування дій, використання інструментів, виконання з перевіркою та контролю безпеки. У контурі, описаному далі, постановка цілей і контроль лишаються за людиною, використання інструментів реалізує сполучний шар командного рядка, а виконання з перевіркою розгорнуто в процеси реалізації, рецензування й аудиту.

Відповідність між етапами традиційного циклу і процесами контуру подає таблиця 1; глибина інтеграції нерівномірна за задумом — найбільша на інженерних етапах від проєктування до супроводу, часткова на формуванні вимог і розгортанні, відсутня лише на моніторингу, який лишається за штатними засобами. Графічно модель показує рисунок 1: етапи утворюють ітеративний цикл; у центрі — людина-оркестратор, що діє через сполучний шар інструментів, спільний для неї та агентів; для кожного етапу зазначено розподіл виконання між агентом і людиною.

Таблиця 1

Інтеграція LLM-агентів в етапи традиційного життєвого циклу

Етап традиційного циклу	Процес контуру	Інтеграція агента
Формування вимог	основні вимоги й обмеження задає інженер; агент фіксує їх і розширює похідними деталізованими вимогами за результатами глибокого аналізу	часткова; основні вимоги й обмеження за людиною
Проєктування	планування (плановий документ у репозиторії) та декомпозиція	повна; затвердження за людиною
Програмування	реалізація: код, тести, збірка, локальні перевірки	повна; перечитування коду інженером
Тестування й верифікація	двоступеневе рецензування з незалежною верифікацією знахідок	повна; вердикти за людиною
Інтеграція та розгортання	конвеєри CI автоматизовані класично; агент ускладнює й оптимізує конвеєри, аналізує журнали збоїв	часткова; релізні рішення за людиною
Супровід	періодичний аудит кодової бази зі структурованими звітами	повна; тріаж за людиною
Моніторинг	штатні системи моніторингу	відсутня

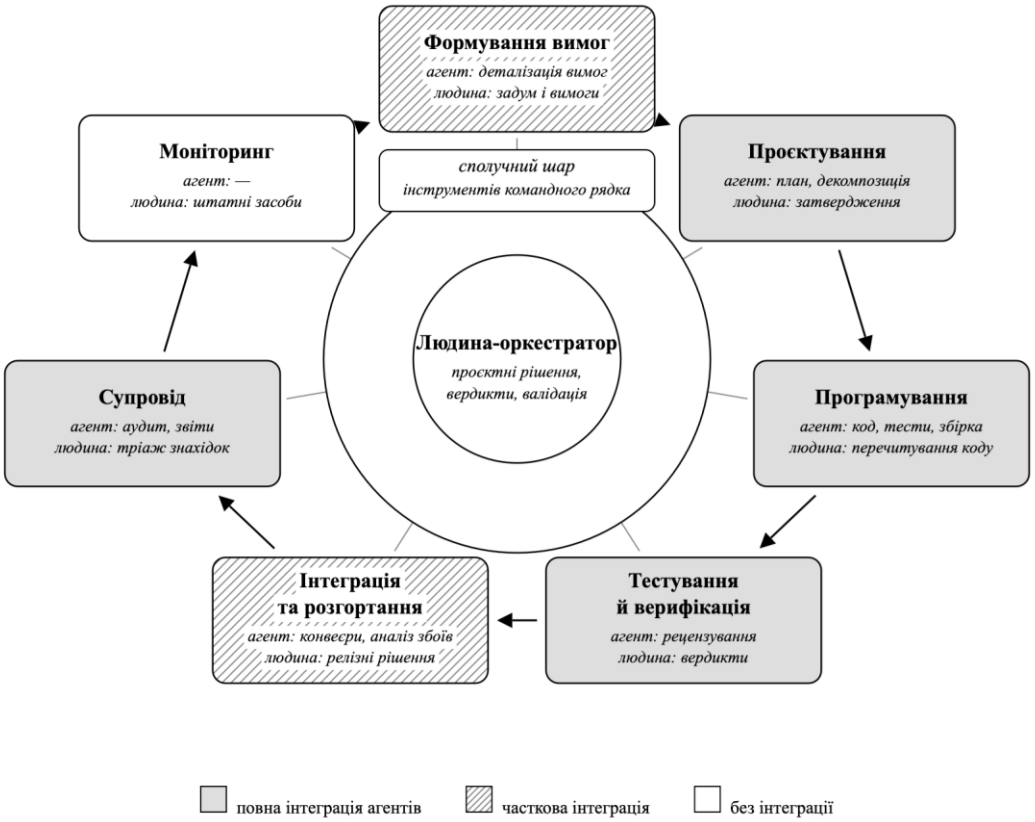


Рисунок 1 — Модель життєвого циклу з інтеграцією LLM-агентів: заливка етапів кодує глибину інтеграції; у кожному етапі — розподіл функцій між агентом і людиною; у центрі — людина-оркестратор і сполучний шар інструментів командного рядка

Планування. Точка інтеграції на проєктувальному етапі — плановий документ. Він народжується в діалозі: інженер формулює задум, основні вимоги та обмеження; агент, маючи доступ до кодової бази, виконує глибокий аналіз, розширює початкові вимоги похідними деталізованими й готує структурований план у форматі markdown, який зберігає в самому репозиторії. План проходить редагування й затвердження людиною, версіонується разом із кодом та актуалізується в ході реалізації — розбіжності між планом і фактичним станом робіт усуває той самий агент. Задачі в системі відстеження формуються з планового контексту: агент викликає інструмент командного рядка поверх API трекера і створює тікети з описами, зібраними з плану. Ручне перенесення контексту між документами й трекером — операція, що раніше забирала помітну частку часу, — з контуру зникає.

Декомпозиція. На тому самому проєктувальному етапі агентові передано й декомпозицію. Роботу розбито на фази за класичними критеріями зв'язності й модульності — з поправкою на можливості агента: класична фаза може містити кілька підфаз, розмір яких узгоджено з обсягом, що його агент надійно утримує в одній керованій сесії; всередині фази агент організовує етапи самостійно. Розбиття і залежності пропонує агент на підставі аналізу кодової бази, людина коригує межі. Оцінювання обсягу задач лишилося класичним — в умовних одиницях складності (story points), без участі агента; при цьому у складність задач закладається використання ШІ-засобів, що дещо знижує номінальні оцінки. Незмінність методики оцінювання суттєва для валідності вимірювань, наведених далі.

Реалізація. Етап програмування інтегровано найглибше: зміни виконує агент у межах затвердженого плану — пише код і тести, збирає проєкт, проганяє локальні перевірки. Обов'язковим елементом етапу є повне перечитування фінального коду інженером: досвід упровадження показав, що відмова від цього кроку призводить до кратного падіння якості коду й дотримання конвенцій. Агентні перевірки наступних етапів доповнюють це перечитування, але не заміняють його.

Рецензування. Верифікаційний етап циклу закриває двоступеневе рецензування. Перший ступінь — самоперевірка до відкриття запиту на злиття: зміни паралельно аналізують кілька спеціалізованих агентів-рецензентів — за коректністю, безпекою, узгодженістю з історією змін репозиторію, надлишковою складністю; їхні знахідки надходять окремому агенту-верифікатору, який перевіряє кожну на відтворюваність із протилежною настановою — спростувати — і відсіює хибні спрацювання; підтверджені знахідки виправляються до відкриття запиту. Схема принципово відрізняється від наївного «попросити модель перевірити власний код»: рецензенти працюють у контекстах, відокремлених від контексту генерації, а верифікація знахідок виділена в окрему фазу.

Другий ступінь — зовнішнє рецензування відкритого запиту людьми та агентним рецензентом у конвеєрі CI. Зауваження обробляє агент на робочій машині автора змін: читає гілку обговорення через інструмент командного рядка, готує виправлення й відповіді; автор валідує їх перед надсиланням. Цикл «зауваження — виправлення — відповідь», який за ручної обробки розтягується через перемикання контексту, стискається до одного проходу валідації.

Розгортання. Цей етап був автоматизований задовго до агентів: конвеєри CI виконують збірку, перевірки та розгортання за усталеними сценаріями. Агентний внесок двоякий: вартість супроводу конвеєрів упала, що дає змогу робити їх складнішими й оптимізованішими; стиснувся й цикл усунення збоїв — агент через інструменти командного рядка читає журнали виконання, локалізує причину відмови та готує виправлення. Релізні рішення й запуск лишаються за людиною.

Аудит кодової бази. Етап супроводу отримав новий регулярний процес — аудит силами агентів. Позаплановий аудит виконується перед публічними релізами; плановий — щомісяця на активних розробках, якщо релізного аудиту не було. Аудит є повним агентним аналізом кодової бази за напрямами: архітектурні відхилення, технічний борг, потокобезпека, безпека, узгодженість конвенцій. Вихід — структурований звіт, у якому кожна знахідка споряджена посиланням на файл і рядок та пріоритетом за впливом; звіт тріажить людина, рішення про виправлення не автоматизоване.

Перебіг типового циклу показує таблиця 2 — релізний аудит SDK для iOS (шість модулів) у березні 2026 року. Первинний аудит дав 38 знахідок; серед критичних — шляхи аварійного завершення застосунку-носія та відсутність обов'язкового маніфесту приватності платформи. Через п'ять днів, після планового виправлення, повторний аудит підтвердив закриття всіх критичних і високопріоритетних знахідок первинного огляду. Суттєвим елементом циклу є перехресна перевірка: незалежний аудит моделлю іншого постачальника (Codex) виявив вісім додаткових знахідок, зокрема дефект застосування конфігурації, пропущений першою моделлю. Сліпі зони різних моделей не збігаються, і другий прогін дешево розширює покриття.

Таблиця 2

Перебіг релізного циклу аудиту SDK для iOS, березень 2026 року

Дата	Подія	Результат
12.03	Первинний аудит (Claude Opus), шість модулів	38 знахідок: 4 критичні, 10 високих, 17 середніх, 7 низьких
13.03	План виправлень, агентна сесія реалізації	закрито критичні й високі знахідки
15.03	Повторний аудит (Claude Opus)	знахідки первинного аудиту закрито; нових: 2 високі, 5 середніх, 3 низькі
15.03	Незалежний аудит (Codex)	8 знахідок: 3 високі, 2 середні, 3 довідкові; частину першою моделлю не виявлено
16.03	Усунення знахідок обох оглядів	усі дієві знахідки закрито
17.03	Ревалідація (Claude Opus)	збірка успішна; 1970 тестів без відмов; у залишку поодинокі низькопріоритетні зауваження

Сполучний шар інструментів. Наскрізним для всіх етапів є єдиний набір інструментів командного рядка, на який спираються перелічені процеси: робота з трекером, із запитами на злиття, з конвеєрами CI (CircleCI, GitHub), збірка, тестування, релізна автоматизація, отримання даних із внутрішніх систем. Принцип шару: ті самі інструменти використовують і люди, і агенти — окремі інтеграції під кожного споживача не створюються. Показова обставина: більшість цих інструментів існувала до появи агентів і використовувалася рідко, бо людям зручніші веб-інтерфейси; агентний контур різко підняв цінність шару — для агента командний рядок є природним інтерфейсом, а кожен виклик залишає відтворюваний слід у журналах сесій. Альтернативний шлях — інтеграція через протокол MCP — був випробуваний і згорнутий на користь інструментів командного рядка (див. розділ обмежень).

Розподіл функцій між агентом і людиною за процесами контуру підсумовує таблиця 3.

Таблиця 3

Розподіл функцій між агентом і людиною за процесами контуру

Процес	Виконує агент	Лишається за людиною
Планування	аналіз кодової бази, деталізація вимог, чернетка та актуалізація плану, формування тікетів із планового контексту	основні вимоги й обмеження, проєктні рішення, затвердження плану
Декомпозиція	пропозиція фаз і залежностей, внутрішнє розбиття фази на етапи	корекція меж фаз, оцінювання обсягу
Реалізація	зміни коду, тести, збірка, локальні прогони	повне перечитування фінального коду, розв'язання неоднозначностей
Рецензування	самоперевірка спеціалізованими рецензентами, верифікація знахідок, обробка зауважень	вердикти щодо знахідок, валідація виправлень і відповідей
Аудит	повний аналіз кодової бази, структурований звіт зі свідченнями	тріаж знахідок, рішення про виправлення

Метрики та методика вимірювання. Вимірювання в промисловій обстановці підпорядковане трьом вимогам: показники мають обчислюватися з наявних артефактів процесу без додаткового навантаження на команду; бути стійкими до неоднорідності потоку задач; явно розрізняти виміряне й експертно оцінене.

З цих міркувань відхилено метрику тривалості циклу задачі (lead time), центральну для класичних методик DORA [5]: потік задач між порівнюваними періодами неоднорідний за складом, і зміна медіанної тривалості відбивала б зміну номенклатури задач не менше, ніж зміну продуктивності. Відхилено й підрахунок зауважень рецензування: він вимірює суворість рецензентів, а не якість змін.

Прийнято чотири групи показників.

Перша — velocity: сумарна оцінка завершених задач в умовних одиницях складності (story points). Дані відновлено з реєстру задач: для кожної з 848 задач проєкту, що фігурували у спринтах команди за 2025–2026 роки, взято оцінку, виконавця та момент завершення — дату вирішення або, за її відсутності, останню зміну статусу на завершальний за журналом змін задачі; кожну задачу враховано рівно один раз (38 % задач побували більш ніж в одному спринті, тому підрахунок за списками спринтів завищував би результат). Основна одиниця агрегації — календарний місяць за датою завершення: сітка спринтів упродовж періоду змінювалася (дво- й тритижневі, навесні 2026 року — з перекриттям), що робить спринтовий показник менш зіставним; спринтовий розріз використано як перевірку стійкості результату. Валідність порівняння спирається на незмінність методики оцінювання: шкала й процедура не змінювалися протягом усього періоду; єдиний задокументований зсув — часткове врахування ШІ-засобів в оцінках, яке знижує їх номінально, — діє проти вимірюваного ефекту, тож оцінка приросту є нижньою межею. Періодизація: базовий період — січень–липень 2025 року, до появи агентних засобів

у процесі; серпень–листопад 2025 року — індивідуальна практика автора, з листопада — пілот на одному продукті; грудень 2025 — лютий 2026 року — перехідний період командного впровадження (навчання, розгортання інфраструктури інструментів), з порівняння виключений разом із внеском ефекту навчання; командний період — березень–липень 2026 року. Неповний серпень 2026 року, на який припало разове масове закриття застарілих задач, виключено.

Друга — кількість одночасно ведених задач на інженера (WIP): скільки задач інженер фактично просуває паралельно. Показник спостерігається на дошці спринта, але інтерпретується експертно, тому наводиться як оцінка, а не вимірювання.

Третя — органічність коду: відповідність нових змін конвенціям наявної кодової бази. Метрики загального призначення таких конвенцій не вловлюють [10], тому показник операціоналізовано через аудитні перевірки гігієни, специфічні для кодової бази: кількість примусових розгортань опціоналів у робочому коді, подавлень правил статичного аналізатора, маркерів незавершеної роботи, відхилень від конвенцій найменування, а також обсяг і результат прогону автоматизованих тестів.

Четверта — показники циклу аудиту: кількість знахідок за пріоритетами, строк усунення, результат ревалідації та розбіжність знахідок між незалежними моделями.

Дизайн дослідження — квазіекспериментальне зіставлення періодів до та після впровадження на реальних проєктах, без контрольної групи; загрози валідності, що з цього випливають, розглянуто в розділі обмежень.

Результати впровадження. Місячна velocity команди зросла вдвічі: медіана — з 42,5 story points на місяць у базовому періоді до 85,0 у командному, середні значення — з 44,4 до 86,3, тобто в 1,95 раза (рисунок 2). Результат стійкий до вибору одиниці агрегації: у спринтовому розрізі медіана зросла з 24,5 до 41,5 story points за спринт, у 1,69 раза; менший множник пояснюється перекриттям спринтів навесні 2026 року та зарахуванням задач, завершених після планового кінця спринта. Динаміка східчаста: фаза індивідуальної практики та пілоту дала проміжне підвищення (медіана 59,3, на 39 % вище бази; найрезультативніший спринт 2025 року — 41,5 story points — припадає на перші два тижні пілоту), перехідний період повернув показник до базового рівня, що відбиває вартість перебудови процесу, а стійке подвоєння настало з березня 2026 року; пікові місяці — 96 і 98 story points.

Декомпозиція приросту: у шести інженерів, активних в обох періодах, сумарний місячний обсяг зріс у 1,70 раза; ще 16 % обсягу командного періоду припадає на членів команди, які раніше в цьому проєкті майже не працювали, — залучення, уможливлене зниженням вартості входу в кодову базу. В окремих виконавців в окремі місяці зафіксовано зростання у 3,6–5,5 раза проти власної медіани базового періоду; ці піки не відтворюються стабільно, залежать від збігу задач добре пророблені області з повністю розгорнутою інфраструктурою інструментів і частково відбивають перерозподіл задач між людьми, тому наводяться як спостереження, а не як відтворюваний ефект.

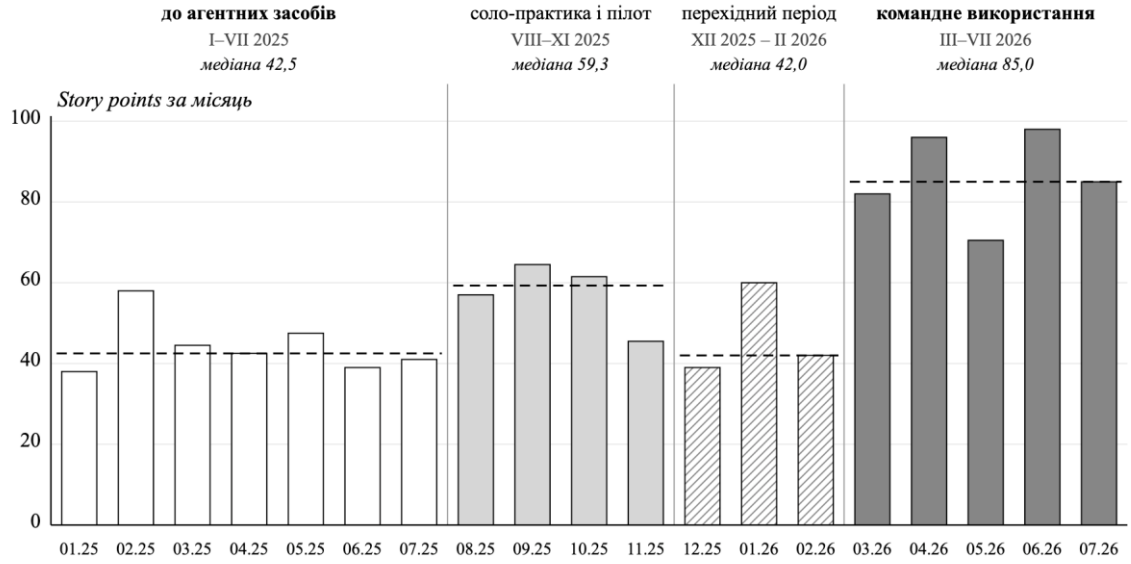


Рисунок 2 — Місячна velocity команди (story points завершених задач на місяць) за чотирма фазами впровадження; пунктиром позначено медіани фаз

Кількість одночасно ведених задач на інженера рівня Senior/Lead зросла з одної–двох до трьох–чотирьох: інженер веде кілька агентних сесій паралельно й перемикається між точками валідації. Стелю паралелізму задає не пропускна здатність агентів, а пропускна здатність людини на вердикти: час на змістовну перевірку плану, знахідки чи зміни агентом не стискається.

Показники циклу аудиту наведено в таблиці 2. Обсяг перевірки, виконаної за п'ять днів — повний аудит шести модулів двома незалежними моделями з повторною ревалідацією, — за ручного виконання

відповідав би тижням роботи рецензентів. Стан кодової бази після циклу за перевірками гігієни: нуль примусових розгортань опціоналів, нуль подавлень статичного аналізатора, нуль маркерів незавершеної роботи в робочому коді, повна відповідність конвенціям найменування, 1970 автоматизованих тестів без відмов.

До агента перейшли й рутинні операції, окреме вимірювання яких не проводилося: оформлення тікетів із планового контексту, актуалізація планових документів, первинна обробка зауважень рецензування, підготовка звітів аудиту. Вивільнення часу тут — якісне спостереження, узгоджене зі зростанням velocity.

Ключове спостереження стосується нерівномірності ефекту: приріст є функцією пророблености області. Стійке подвоєння на рівні команди відповідає повному розгортанню контуру на пілотних продуктах; піки в окремих виконавців — збігу задач добре пророблені області з повною інфраструктурою інструментів командного рядка й усталеними конвенціями. Звідси випливає, що визначальним ресурсом продуктивності агентного контуру є не засіб генерації, а накопичені проєктні знання та інфраструктура інструментів довкола нього; систематичне накопичення таких знань і оцінювання їх достовірності становить окрему наукову задачу, що виходить за межі цієї статті.

Обговорення й обмеження. Межі застосовності контуру виявилися стабільними протягом усього періоду спостереження. Найменший ефект — у роботі з візуальними інтерфейсами: цикл зворотного зв'язку замкнений через людину, що перевіряє результат на пристрої, і агент його не стискає. Великі наскрізні зміни через багато модулів чи репозиторіїв деградують за накопичення обмежень і обсягу контексту, що узгоджується з експериментальними спостереженнями деградації агентів за зростання кількості одночасних вимог [2]; такі зміни веде людина, з агентною допомогою на окремих ділянках. Третя межа — міжпродуктова взаємодія: агент, працюючи в репозиторії одного продукту, не має знання про суміжні продукти, і для задач на стику довелося будувати окремі контури й інструменти.

Показовими є відкати — елементи, введені в контур і згодом виведені з нього. Інтеграцію через протокол MCP замінено інструментами командного рядка: за однакової функціональності інтеграція CLI у різнорідну наявну інфраструктуру виявилася простішою і придатною одразу для обох споживачів — людини та агента. Інструмент аналізу залежностей коду GitNexus, доступний агентові й описаний у настановах, агентом систематично не використовувався — приклад ширшого режиму, за якого доступне знання не застосовується; систематичне дослідження цього режиму виходить за межі статті.

Загрози валідності. Дизайн без контрольної групи не відділяє ефекту контуру від паралельних процесів, насамперед від зростання якості самих моделей за період спостереження; результати слід читати як ефект зв'язки «моделі + контур». Кейс обмежений однією організацією й малою командою; контур будував і найінтенсивніше застосовував автор, тож перенесення на команди без внутрішнього носія практики не досліджене. Потік задач неоднорідний, і, попри незмінність методики оцінювання, можливий неспостережений дрейф шкали story points; задокументований компонент дрейфу спрямований на заниження оцінок, тобто діє консервативно. Особисті множники зростання частково відбивають перерозподіл задач між людьми та продуктами, а не лише продуктивність. Нарешті, можливий мимовільний відбір задач під агентні засоби, хоча політика команди його не передбачала.

Висновки та пропозиції. LLM-агенти формують нову парадигму автоматизації життєвого циклу програмного забезпечення, за якої штучний інтелект переходить від ролі пасивного генератора коду до ролі активного учасника процесу розробки. Основними напрямами застосування є автоматизація аналізу вимог, архітектурного проєктування, програмування, тестування, налагодження, управління проєктами, аналізу безпеки, документування та моніторингу.

Найперспективнішою видається багаторівнева мультиагентна архітектура, в якій спеціалізовані LLM-агенти взаємодіють між собою, використовують спільну пам'ять і зовнішні інструменти та проходять цикли незалежної перевірки, — саме за цією схемою в описаному контурі побудовано рецензування й аудит. У перспективі такий підхід трансформує традиційну модель розподілу функцій: людина визначає цілі, політики, обмеження та критерії якості, а LLM-агенти забезпечують адаптивне планування, виконання, перевірку та оптимізацію значної частини процесів життєвого циклу програмного забезпечення.

Основні результати роботи такі.

1. Описано промисловий контур автоматизації процесів SDLC на основі LLM-агентів — інтеграцію агентів в окремі етапи традиційного життєвого циклу, — що охоплює планування з версіонованими плановими документами, декомпозицію, двоступеневе рецензування з незалежною верифікацією знахідок, періодичний аудит кодової бази та шар інструментів командного рядка як спільний інтерфейс людини й агента; наскрізні принципи — валідація людиною всіх дій з незворотними наслідками та повне перечитування фінального коду інженером.

2. За дванадцять місяців експлуатації, з яких вісім командної, місячна velocity команди зросла вдвічі (медіана — з 42,5 до 85,0 story points на місяць; у спринтовому розрізі — у 1,7 раза) за незмінної методики оцінювання задач; кількість одночасно ведених задач на інженера зросла з одної–двох до трьох–чотирьох; релізний цикл аудиту — 38 знахідок, п'ять днів, перехресна перевірка моделями двох постачальників — продемонстрував глибину перевірки, недосяжну за ручного виконання в порівнянний строк.

3. Приріст нерівномірний і визначається пророблені стю області: накопиченими інструкціями, інфраструктурою інструментів, усталеністю конвенцій — від стійкого подвоєння на рівні команди до піків у 3,6–5,5 раза в окремих виконавців у місяці

збігу цих чинників. Визначальним ресурсом продуктивності є, отже, не засіб генерації, а накопичені проєктні знання й організація процесу довкола нього.

4. Межі застосовності: візуальні інтерфейси, великі наскрізні зміни, міжпродуктова взаємодія. Стелю паралелізму задає пропускна здатність людини на вердикти; роль людини зміщується від виконавця до оркестратора, за яким лишаються проєктні рішення та вердикти.

5. Напрямом подальших досліджень є систематичне накопичення проєктних знань, що визначають пророблені область, та оцінювання їх достовірності й чинності: саме цей ресурс, як показує кейс, обмежує досяжний приріст продуктивності агентного контуру.

References

1. Becker J., Rush N., Barnes E., Rein D. Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity. 2025. arXiv:2507.09089.

2. Constraint Decay: The Fragility of LLM Agents in Backend Code Generation. 2026. arXiv:2605.06445.

3. Cui Z., Demirer M., Jaffe S., Musolff L., Peng S., Salz T. The Effects of Generative AI on High-Skilled Work: Evidence from Three Field Experiments with Software Developers. *Management Science*. 2025.

4. DORA. State of AI-assisted Software Development 2025: research report. Google Cloud, 2025.

5. Forsgren N., Humble J., Kim G. Accelerate: The Science of Lean Software and DevOps. Portland: IT Revolution Press, 2018. 288 p.

6. Harding B. AI Copilot Code Quality: 2025 Data Suggests 4x Growth in Code Clones: research report. GitClear, 2025.

7. He J., Treude C., Lo D. LLM-Based Multi-Agent Systems for Software Engineering: Literature Review, Vision and the Road Ahead. *ACM Transactions on Software Engineering and Methodology*. 2025. Vol. 34, № 5. P. 1–30.

8. Peng S., Kalliamvakou E., Cihon P., Demirer M. The Impact of AI on Developer Productivity: Evidence from GitHub Copilot. 2023. arXiv:2302.06590.

9. Sommerville I. Software Engineering. 10th ed. Boston: Pearson, 2016. 816 p.

10. Vasilevski K., Dong X., Rombaut B. та ін. Beyond Correctness: Enhancing Architectural Reasoning in Code LLMs via Scalable Labeling with Agentic Judgment. 2026. arXiv:2606.14948.

11. Wang H., Poskitt C. M., Sun J. AgentSpec: Customizable Runtime Enforcement for Safe and Reliable LLM Agents. *Proc. 48th IEEE/ACM International Conference on Software Engineering (ICSE '26)*. Rio de Janeiro, 2026. arXiv:2503.18666.

12. Ziegler A., Kalliamvakou E., Li X. A. та ін. Measuring GitHub Copilot's Impact on Productivity. *Communications of the ACM*. 2024. Vol. 67, № 3. P. 54–63.

13. ДСТУ ISO/IEC/IEEE 12207:2018. Інженерія систем і програмних засобів. Процеси життєвого циклу програмних засобів (ISO/IEC/IEEE 12207:2017, IDT). Чинний від 2018-08-15. Київ: ДП «УкрНДНЦ», 2018.

No 195 (2026)

Sciences of Europe
(Praha, Czech Republic)

ISSN 3162-2364
The journal is registered and published in Czech Republic.
Articles in all spheres of sciences are published in the journal.

Journal is published in Czech, English, Polish, Russian, Chinese, German and French, Ukrainian.
Articles are accepted each month.
Frequency: 24 issues per year.
Format - A4
All articles are reviewed
Free access to the electronic version of journal

Edition of journal does not carry responsibility for the materials published in a journal.
Sending the article to the editorial the author confirms it's uniqueness and takes full responsibility for possible consequences for breaking copyright laws.

Chief editor: Petr Bohacek
Managing editor: Michal Hudecek

- Jiří Pospíšil (Organic and Medicinal Chemistry) Zentiva
- Jaroslav Fähnrich (Organic Chemistry) Institute of Organic Chemistry and Biochemistry Academy of Sciences of the Czech Republic
- Rasa Boháček – Ph.D. člen Česká zemědělská univerzita v Praze
- Naumov Jaroslav S., MD, Ph.D., assistant professor of history of medicine and the social sciences and humanities. (Kiev, Ukraine)
- Viktor Pour – Ph.D. člen Univerzita Pardubice
- Petrenko Svyatoslav, PhD in geography, lecturer in social and economic geography. (Kharkov, Ukraine)
- Karel Schwaninger – Ph.D. člen Vysoká škola báňská – Technická univerzita Ostrava
- Václav Pittner -Ph.D. člen Technická univerzita v Liberci
- Dudnik Oleg Arturovich, Doctor of Physical and Mathematical Sciences, Professor, Department of Physical and Mathematical management methods. (Chernivtsi, Ukraine)
- Konovalov Artem Nikolaevich, Doctor of Psychology, Professor, Chair of General Psychology and Pedagogy. (Minsk, Belarus)

«Sciences of Europe» -
Editorial office: Křižíkova 384/101 Karlín, 186 00 Praha

E-mail: info@european-science.org
Web: www.european-science.org